Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Modelos de minería de datos: random forest y adaboost, para identificar los factores asociados al uso de las TIC (internet, telefonía Fija y televisión de paga) en los hogares del Perú. 2014

TESIS

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Jorge Brian ALARCÓN FLORES

ASESOR

María Estela PONCE ARUNERI

Lima, Perú

2017

Referencia bibliográfica

Alarcón, J. (2017). *Modelos de minería de datos: random forest y adaboost, para identificar los factores asociados al uso de las TIC (internet, telefonía Fija y televisión de paga) en los hogares del Perú. 2014*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
(Universidad del Perú, DECANA DE AMÉRICA)
FACULTAD DE CIENCIAS MATEMÁTICAS

113

ESCUELA PROFESIONAL DE ESTADÍSTICA

ACTA DE SUSTENTACIÓN DE TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA

En la Ciudad Universitaria, Facultad de Ciencias Matemáticas, siendo las 15:30 horas del día 18 de Mayo del año 2017 se reunieron los docentes designados como miembros del Jurado:

Mg. Emma Norma Cambillo Moyano	(Presidenta)
Mg. Richard Fernando Fernández Vásquez	(Miembro)
Mg. María Estela Ponce Aruneri	(Miembro Asesor)

Para la sustentación de la Tesis intitulada "**MODELOS DE MINERÍA DE DATOS: RANDOM FOREST Y ADABOOST, PARA IDENTIFICAR LOS FACTORES ASOCIADOS AL USO DE LAS TIC (INTERNET, TELEFONÍA FIJA Y TELEVISIÓN DE PAGA) EN LOS HOGARES DEL PERÚ. 2014**", presentada por el Bachiller Jorge Brian ALARCÓN FLORES, para obtener el Título Profesional de Licenciado en Estadística.

Luego de la exposición de la Tesis, la Presidenta invitó al expositor a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del jurado, el expositor mereció la aprobación.....BUENO....., con un calificativo promedio de ...DIECISEIS (16).....(letras y números).

A continuación los miembros del jurado, dan manifiesto que el participante Bachiller Jorge Brian ALARCÓN FLORES, en virtud de haber aprobado la sustentación de su tesis, será propuesto para que se le otorgue el Título Profesional de Licenciado en Estadística.

Siendo las 16:30 horas, se levantó la Sesión, firmando para constancia la presente Acta en tres (3) copias originales.

Mg. Richard Fernando Fernández Vásquez
Miembro

Mg. María Estela Ponce Aruneri
Miembro Asesor

30 ENE 2018

Mg. Emma Norma Cambillo Moyano
Presidenta

FICHA CATALOGRÁFICA

ALARCON FLORES, JORGE BRIAN

Modelos de minería de datos: Random Forest y Adaboost, para identificar los factores asociados al uso de las TIC (Internet, Telefonía Fija y Televisión de Paga) en los hogares del Perú, 2014, (Lima) 2017.

vii, 97 p., 29.7 cm, (UNMSM, Licenciado, Estadística, 2017).

Tesis, Universidad Nacional Mayor de San Marcos,

Facultad de Ciencias Matemáticas 1. Estadística

I. UNMSM/FACULTAD DE CIENCIAS MATEMÁTICAS

DEDICATORIA

Por iluminarme y protegerme en cada momento de mi vida:

A Dios.

Por tu gran amor y lucha constante, este logro es para ti:

Mamá.

Por creer en mí y alentarme siempre a seguir adelante, a ustedes:

Hermanita, papá y mamá Julia.

AGRADECIMIENTO

Por su guía profesional y apoyo incondicional, a mi asesora:

Mg. María Estela Ponce Aruneri.

Y a todas aquellas personas que estuvieron presentes siempre en este largo camino universitario.

RESUMEN

MODELOS DE MINERÍA DE DATOS: RANDOM FOREST Y ADABOOST, PARA IDENTIFICAR LOS FACTORES ASOCIADOS AL USO DE LAS TIC (INTERNET, TELEFONÍA FIJA Y TELEVISIÓN DE PAGA) DE LOS HOGARES EN PERÚ, 2014.

Br. Jorge Brian Alarcón Flores

MAYO– 2017

ASESORA : Mg. María Estela Ponce Aruneri
TÍTULO OBTENIDO: Licenciado en Estadística

La sociedad hoy en día se encuentra viviendo una etapa de constantes cambios, debidos en gran medida a la introducción de nuevas tecnologías en la vida cotidiana, diversos líderes mundiales afirman que el uso de las Tecnologías de la Información y Comunicación (TIC) juegan un rol fundamental en el desarrollo de las naciones. El objetivo de este estudio es identificar los factores asociados al uso de las TIC en los hogares del Perú, y que deberán considerarse en las políticas sociales y económicas ligadas a la accesibilidad y manejo de tecnologías en nuestro país, como vía de progreso social y desarrollo nacional.

Se considera como TIC para la presente tesis, el acceso a los servicios de internet, telefonía fija o tv de paga. Con la base de datos de la Encuesta Residencial de Servicios de Telecomunicaciones (ERESTEL) realizada por OSIPTEL, se aplicaron los modelos de minería de datos de clasificación supervisada, random forest y adaboost, para identificar los factores asociados al uso de estas tecnologías en los hogares del Perú, obteniendo como resultados que son la lengua materna, el nivel de estudio del jefe del hogar y el área sociodemográfica donde se encuentra ubicada la vivienda, como los factores más importantes asociados al uso de las TIC en los hogares del Perú en el 2014. Finalmente se determinó los patrones de consumo de las TIC en los hogares peruanos, utilizando el modelo árboles de clasificación.

Palabras clave: *random forest, adaboost, árbol de clasificación, TIC.*

ABSTRACT

DATA MINING MODELS: RANDOM FOREST AND ADABOOST, TO IDENTIFY FACTORS ASSOCIATED WITH THE USE OF TIC (INTERNET, FIXED TELEVISION AND PAY TV) OF HOUSEHOLDS IN PERU, 2014.

Br. Jorge Brian Alarcón Flores

MAY- 2017

ADVISOR : Mg. María Estela Ponce Aruneri

DEGREE OBTAINED: Bachelor's degree in Statistics

Society today is experiencing a period of constant change, due largely to the introduction of new technologies in everyday life, several world leaders claim that the use of Information and Communication Technologies (ICT) play a Fundamental role in the development of nations. The objective of this study is to identify the factors associated with the use of ICTs in Peruvian households, which should be considered in the social and economic policies linked to the accessibility and management of technologies in our country, as a way of social progress and development national.

It is considered as ICT for this thesis, access to internet services, fixed telephony or pay tv. With the OSELTEL database of the Residential Survey of Telecommunications Services (ERESTEL), supervised classification data mining models, random forest and adaboost were applied to identify factors associated with the use of these technologies in households Of Peru, obtaining as results that are the mother tongue, the level of study of the head of the household and the sociodemographic area where the house is located, as the most important factors associated with the use of the ICT in Peruvian households in 2014. Finally, we determined the consumption patterns of ICTs in Peruvian households, using the classification trees model.

Keywords: *random forest, adaboost, classification tree, ICT*

INTRODUCCIÓN

Los grandes avances que viene teniendo la tecnología en el mundo, la sitúan actualmente como una de las temáticas de estudio de mayor relevancia en la sociedad y en la vida cotidiana de los hogares. Pero ¿Cuál es la situación real de los hogares peruanos respecto al uso de las tecnologías?, según el ranking del Informe Global de Tecnologías de la Información 2016, Perú se ubica en el puesto 90 en el uso de las TIC de 139 economías mundiales (Sociedad Nacional de Industrias, 2016), situación que preocupa, sabiendo que hoy en día el acceso a tecnologías, como el Internet, son consideradas como fuentes económicas y de desarrollo social muy importante. Es por eso, que este trabajo de investigación tiene como objetivo identificar los factores asociados al uso de las TIC (Internet, Telefonía fija y Televisión de paga) en los hogares del Perú en 2014.

Para identificar los factores asociados al uso de las TIC en los hogares del Perú, se utilizaron las técnicas de minería de datos: *random forest* y *adaboost*, consideradas como dos de las técnicas más precisas y eficientes que existen actualmente. En el caso de *random forest*, miembros destacados de la comunidad *Data Science Central*, como Michael Walker, Director General de *Rose Business Technologies* afirman: "El modelo *random forest* es uno de los mejores entre los modelos de clasificación, capaz de clasificar grandes cantidades de datos con exactitud". Respecto al modelo *adaboost*, en 2003, sus creadores Yoav Freund y Robert Schapire ganan el prestigioso Premio Godel, por su importante contribución en el área de las ciencias de la computación con el algoritmo, el cual es considerado

como uno de los modelos más versátiles y rápidos del *machine learning* (Emer, 2012).

Adicionalmente se determinó los patrones de consumo de las TIC en los hogares peruanos, utilizando el modelo árboles de clasificación.

Este trabajo de investigación comprende 4 capítulos, en el Capítulo I se presenta el planteamiento del problema, los objetivos, justificación e hipótesis, útiles para una mejor comprensión del contexto de la investigación.

El capítulo II contiene el marco teórico, donde se presentan antecedentes nacionales e internacionales de la investigación, además de las bases teóricas de los modelos de minería de datos utilizados.

El capítulo III abarca la parte metodológica de la tesis, desde presentación e identificación de variables hasta los aspectos relacionados al instrumento y software utilizados.

Finalmente, en el capítulo IV se presentan los resultados obtenidos en la identificación de factores asociados al uso de las TIC en los hogares del Perú con los modelos *random forest* y *adaboost*; y de los patrones de consumo de las TIC mediante árboles de clasificación.

INDICE

I.PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1.Situación Problemática

El mundo entero se encuentra viviendo desde hace unos años atrás constantes cambios, debidos en gran medida a la introducción de nuevas tecnologías en el quehacer de la vida cotidiana. En el año 2003, los más importantes líderes mundiales declararon*: "*Somos plenamente conscientes de que las ventajas de la revolución de la tecnología de la información están en la actualidad desigualmente distribuidas entre los países desarrollados y en desarrollo, así como dentro de las sociedades" (Cumbre Mundial de la Sociedad de la Información, 2003).

Han transcurrido trece años desde aquel entonces y si bien han mejorado los indicadores de accesibilidad al uso de las TIC en los hogares peruanos, aún siguen existiendo grandes brechas entre los hogares que utilizan y no estas tecnologías; según el ranking del Informe Global de Tecnologías de la Información 2016, Perú se mantiene en el puesto 90 en el uso de las TIC de 139 economías mundiales (Sociedad Nacional de Industrias, 2016).

Para muchos estudiosos de las nuevas tecnologías a nivel mundial, las TIC y sobre todo Internet ha sido una de más grandes innovaciones del hombre en la historia, tal como lo menciona Castells (2014), quien afirma que *"*Internet es la

tecnología decisiva de la era de la información, del mismo modo que el motor eléctrico fue el vector de la transformación tecnológica durante la era industrial", concepto que cada vez se aproxima más a lo que vivimos en la actualidad, pues hoy en día Internet ya es mucho más que una herramienta de comunicación, interacción, o "la autopista de la información", como le gustaba decir a algunos expertos en el mundo digital durante la década del '90, sino que hoy en día el acceso y uso de las TIC también se ha convertido en una fuente económica y de desarrollo social muy importante.

Es por eso, que se busca con esta investigación, conocer cuáles son los factores asociados al uso de las TIC en los hogares del Perú.

1.2. Formulación del Problema

De los argumentos expuestos anteriormente surgen muchas interrogantes de carácter investigativo en lo referente a los factores que están detrás del uso de las TIC en los hogares del Perú y las técnicas de minería de datos que se utilizarán para obtener los resultados de esta investigación. Ante esto se formulan las siguientes preguntas de investigación:

1.2.1. Problema general

¿Cuáles son los factores asociados al uso de las TIC (Internet, telefonía fija y TV de paga) en los hogares del Perú obtenidos mediante los modelos de minería de datos *random forest* y *adaboost*?

1.2.2. Problemas específicos:

- ¿Cuáles son los factores asociados al uso del internet en los hogares del Perú obtenidos mediante los modelos de minería de datos *random forest* y *adaboost*?
- ¿Cuáles son los factores asociados al uso de la telefonía fija en los hogares del Perú obtenidos mediante los modelos de minería de datos *random forest* y *adaboost*?
- ¿Cuáles son los factores asociados al uso de la tv de paga en los hogares del Perú obtenidos mediante los modelos de minería de datos *random forest* y *adaboost*?
- ¿Cuáles son los patrones de consumo de las TIC en los hogares del Perú obtenidos mediante el modelo de árbol de clasificación?

La identificación de factores asociados al uso de las TIC, internet, telefonía fija y tv de paga se realizó mediante los modelos de minería de datos *random forest* y *adaboost*, ambas desarrolladas con el software estadístico R Studio versión 1.0.136, y se compararon los modelos de clasificación mediante dos métricas distintas (indicadores estadísticos): sensibilidad y curva de ROC. Para conocer los patrones de consumo de las TIC en los hogares del Perú, se utilizó el modelo de árboles de clasificación, desarrollado en el software SPSS, y validado mediante las métricas de matriz de error y exactitud.

1.3. Justificación de la investigación

Como bien se ha mencionado, hoy en día el acceso a la información inmediata puede ofrecer claras ventajas competitivas, tanto en el ámbito económico,

profesional, social, entre otras; además de contribuir directamente en el desarrollo de una nación; por lo que es importante identificar los factores que se encuentran asociados al uso de las TIC en hogares peruanos. Pero ¿Cómo se traducen estas teorías en términos reales? Muy simple: tomando como ejemplo el caso de Internet, que permite que las Pymes logren competitividad y profundicen su penetración en el mercado global. En los seis mercados emergentes más importantes del mundo, se estima que un 1.3% de los puestos laborales ya están relacionados con actividades online, según informe de la consultora (Mc Kinsey & Company, 2012).

Si bien en la actualidad, entidades públicas, como el Instituto Nacional de Estadística e Informática (INEI) y el Organismo Supervisor de Inversión Privada en Telecomunicaciones (OSIPTEL), presentan informes trimestrales y anuales sobre la evolución del acceso a estas tecnologías en los hogares del Perú, no muestran detalle sobre los factores que más influyen para que éstos hogares hagan uso de las TIC, por lo cual, esta investigación, busca contribuir con dicha información, lo cual además, se encuentra vinculado con el segundo objetivo estratégico del Plan Estratégico Institucional (PEI) de nuestra universidad, que se refiere al desarrollo humano y sostenible, proporcionando así información confiable para la toma de decisiones de las instituciones gubernamentales, y sirviendo como guía a estudiantes e investigadores interesados en realizar investigación sobre minería de datos aplicado a ámbitos sociales.

Por el lado científico, las técnicas y métodos de aprendizaje de modelos comprensibles y proposicionales, propuestos en la presente investigación son

aplicados a nivel internacional en diversas áreas de la ciencia y tecnología. En la actualidad, grandes empresas tales como Facebook, Google y Amazon, tienen áreas dedicadas a minería de datos; el caso más sobresaliente de estas 3 puede ser el de Amazon, empresa líder mundial en comercio electrónico, que ha llevado la minería de datos a todas sus áreas, para hacer cosas tan interesantes como predicción de la demanda de sus productos, fijar precios idóneos a sus productos, hacer recomendaciones personalizadas, optimizar las rutas de distribución o inclusive detectar fraude (Lange, 2016). Si bien estos modelos ya tienen múltiples aplicaciones en temas económicos, negocios, salud, entre otros; no han sido profundizados para estudiar el uso de las TIC en los hogares peruanos, lo cual demuestra la originalidad y valor agregado de esta tesis, frente a otras investigaciones realizadas en el campo de la minería de datos en el Perú.

1.4. Objetivos

1.4.1. Objetivo general

- Identificar los factores asociados al uso de las TIC (Internet, Telefonía fija y Televisión de paga) en los hogares del Perú en 2014 mediante los modelos de minería de datos *random forest* y *adaboost*.

1.4.2. Objetivos específicos

- Determinar los factores asociados al uso del internet en los hogares del Perú en 2014 mediante los modelos de minería de datos *random forest* y *adaboost*.

- Identificar los factores asociados al uso de televisión de paga en los hogares del Perú en 2014 mediante los modelos de minería de datos *random forest* y *adaboost*.

- Describir los factores asociados al uso de la telefonía fija en los hogares del Perú en 2014 mediante los modelos de minería de datos *random forest* y *adaboost*.
- Identificar los patrones de consumo de las TIC en los hogares del Perú en 2014 mediante el modelo de árbol de clasificación.

1.5. Hipótesis

1.5.1. Hipótesis general

- Los factores asociados en el uso de las TIC (internet, telefonía fija y televisión de paga) en los hogares del Perú en 2014 obtenidos mediante el modelo *adaboost*, validado con los indicadores de sensibilidad y ROC, son la lengua materna y nivel de estudios del jefe del hogar, además del área sociodemográfica donde se ubica la vivienda.

1.5.2. Hipótesis específicas

- Los factores asociados al uso de internet en los hogares del Perú en 2014 obtenidos mediante el modelo *adaboost*, validado con los indicadores de sensibilidad y ROC, son la lengua materna y el nivel de estudios del jefe del hogar.

- Los factores asociados al uso de la tv de paga en los hogares del Perú en 2014 obtenidos mediante el modelo *adaboost*, validado con los indicadores de sensibilidad y ROC, son el nivel de estudios y la lengua materna del jefe del hogar, y el área sociodemográfica donde se ubica la vivienda.

- Los factores asociados al uso de la telefonía fija en los hogares del Perú en 2014 obtenidos mediante el modelo *adaboost*, validado con los indicadores de sensibilidad y ROC, son la lengua materna y nivel de estudios del jefe del hogar, y el nivel de pobreza del hogar.

- Los hogares del Perú que más hacen uso de las TIC obtenidos mediante el modelo de árbol de clasificación, validado con los indicadores de sensibilidad y ROC, son aquellos donde los jefes de hogar hablan el idioma castellano, que poseen niveles de estudios superiores (técnico o universitarios) y cuyas viviendas se encuentran ubicadas en áreas urbanas del país.

II. MARCO TEÓRICO

2.1. Antecedentes de investigación

La importancia que vienen adquiriendo las TIC en los últimos años ha sido motivo de estudio y análisis de manera rigurosa por diversos autores, sin embargo, los estudios realizados a nivel nacional como internacional, sólo utilizan métodos y técnicas descriptivas univariadas.

No existen publicaciones que muestren aplicaciones con modelos y técnicas de minería de datos de clasificación supervisada, para identificar los factores asociados del uso de las TIC en los hogares del Perú, usando *random forest* o *adaboost*.

Dentro de las investigaciones sobre factores asociados uso de las TIC en los hogares se encontraron las siguientes:

2.1.1. Antecedentes a nivel nacional

- **(Barrantes Roxana, 2005)**, desarrolló en el INSTITUTO DE ESTUDIOS PERUANOS, LIMA-PERÚ; el trabajo de investigación titulado: "ANÁLISIS DE LA DEMANDA POR TICS: ¿QUÉ ES Y CÓMO MEDIR LA POBREZA DIGITAL?" y que tuvo como

objetivo realizar estimaciones sobre el nivel de pobreza digital en América Latina y el Caribe; los resultados más importantes de la investigación fueron la diferencia en la demanda de las TICS entre los pobres extremos y los no pobres son muy claras y atribuibles a los factores que tienden a explicar la pobreza económica como nivel de educación, ingresos, actividad económica principal, condición urbana, etc.

- **(Instituto Nacional de Estadística e Informática, 2016),** organismo oficial peruano que presenta de manera trimestral el informe: "Estadísticas de las Tecnologías de Información y Comunicación en los hogares", documento elaborado en base a los resultados obtenidos de la Encuesta Nacional de Hogares (ENAHO), publicado desde el 2005. Entre los resultados más importantes de la última actualización del 2016, se observa que el uso de las TIC en los hogares se ha incrementado en mayor medida en los hogares con mayor nivel de educación. Además, el tamaño del hogar y la vivienda siguen siendo algunos de los factores más determinantes en el uso de las TIC en los hogares peruanos.

2.1.2. Antecedentes a nivel internacional

- **(Observatorio para la Sociedad de la Información en Latinoamérica y el Caribe, 2008)**; con el trabajo de investigación titulado: "CARACTERÍSTICAS DE LOS HOGARES CON TIC EN

AMÉRICA LATINA Y EL CARIBE", mostro que lo que más afecta a la disponibilidad de TIC en los hogares es el ingreso, seguido por los años de estudio del jefe del hogar. Por otra parte, la disponibilidad de electricidad y la categoría de la ocupación del jefe del hogar parecen tener un impacto menor. Además, algunos de los resultados de la investigación muestran que Perú es uno de los países que poseen niveles relativamente bajos de penetración de las TIC en los hogares.

Dentro de investigaciones con aplicación de *random forest* y *adaboost* se encontraron sólo a nivel internacional:

- **(Pereira Nicole, 2014),** desarrolló en la UNIVERSIDAD DE CHILE, el trabajo de investigación titulado: "IDENTIFICACIÓN DE CLIENTES CON PATRONES DE CONSUMO ELÉCTRICO FRAUDULENTO" en dicha investigación se utiliza el modelo *random forest* para identificar aquellos consumidores que poseen una alta propensión al hurto de electricidad. Para lo cual se utilizó la información histórica disponible de los clientes desde enero de 2012 a marzo de 2014, tales como consumo mensual, inspecciones previas, cortes de suministro, entre otras fuentes.

- **(Wu Hong, 2011),** desarrolló en la Universidad de Arizona, USA; el trabajo de investigación titulado: "OFFLINE AND ONLINE ADABOOST FOR DETECTING ANATOMIC STRUCTURES", en dicha investigación se utiliza el modelo *adaboost*, en la detección de

estructuras anatómicas del ser humano, tales como el tronco pulmonar y el arco aórtico, para el diseño de un sistema que pueda detectar de manera temprana la embolia pulmonar.

- **(Hop Walter, 2013),** desarrolló en la Universidad de Rotterdam, HOLANDA; el trabajo de investigación titulado: "WEB-SHOP ORDER PREDICTION USING MACHINE LEARNING", el objetivo de dicha investigación es utilizar el modelo *random forest*, para ofrecer opciones de compras personalizadas a los usuarios de una página web de venta electrónica, a partir de compras que hayan realizado con anterioridad a través del mismo portal web.

- **(Radha Krishna, 2009),** desarrolló en la Universidad de Dublín, IRLANDA; el trabajo de investigación titulado: "SPEEDING UP ADABOOST OBJECT DETECTION WITH MOTION SEGMENTATION AND HAAR FEATURE ACCELERATION", en dicha investigación se utiliza el modelo *adaboost*, para el reconocimiento de rostros y objetos, a partir de los videos obtenidos de cámaras de vigilancia.

2.2. Bases Teóricas

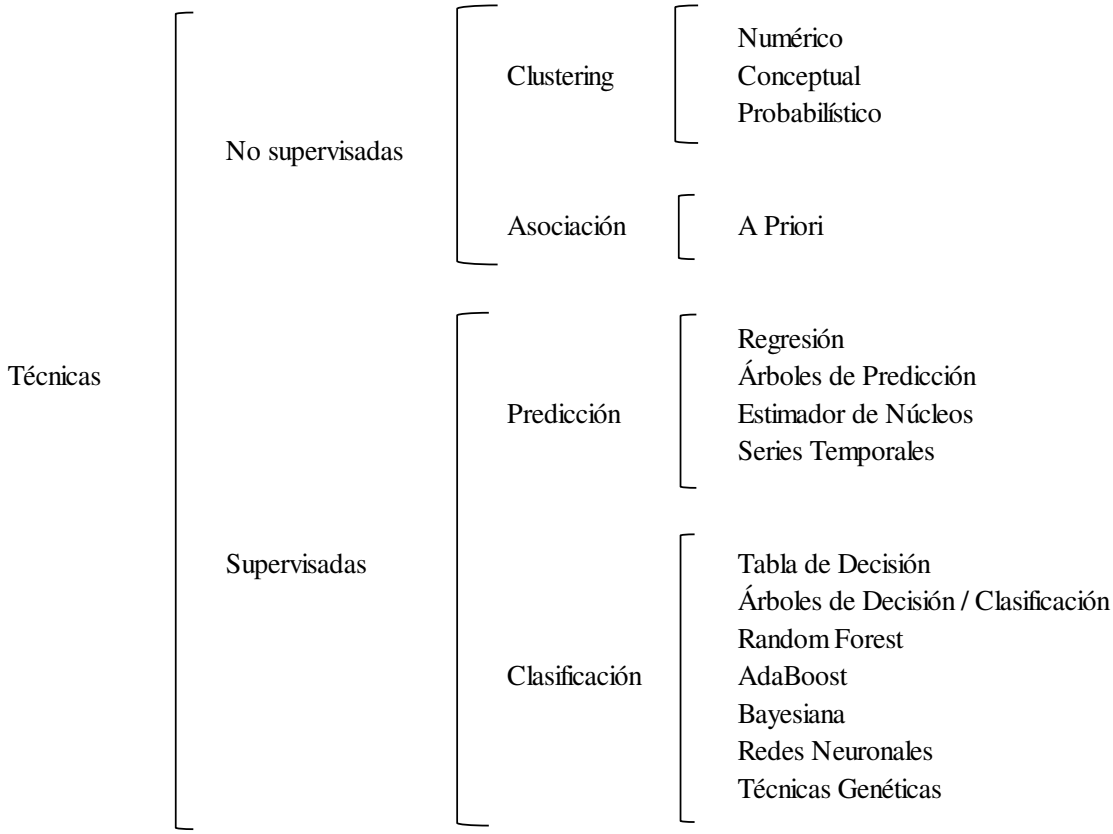
2.2.1. Minería de datos

"La minería de datos es el análisis de conjuntos de datos observacionales (principalmente en grandes cantidades) para el descubrimiento de patrones y para resumir los datos de formas novedosas que sean comprensibles y útiles para el usuario de los datos" (Hand, Mannilla & Smith, 2001, p.18).

Los modelos de minería de datos utilizan técnicas de modelaje estadístico y algoritmos de ciencias de la computación para su desarrollo, esto último, es lo que lo diferencia principalmente del análisis mediante modelos estadísticos, los cuales "utilizan recursos matemáticos para organizar y resumir una gran cantidad de datos obtenidos de la realidad, con el fin de poder tomar decisiones para la solución de problemas, y que pueden ser aplicados de manera descriptiva y de manera inferencial" (Cazau, 2006, p.4). El complemento de la parte de algoritmos computacionales al desarrollo de los modelos estadísticos ha generado que los modelos de minería de datos puedan hoy en día, analizar grandes cantidades de datos, en tiempos de ejecución muchos más reducidos a los que se obtenían con los modelos predecesores de la minería de datos. La descripción del proceso de la utilización de minería de datos, desde la parte de selección, limpieza e integración de datos; hasta la parte del análisis en sí y validación del modelo mediante el uso de softwares de programación recibe el nombre de metamodelo (Espinoza, 2014). Dentro de sus principales aplicaciones, se encuentran investigaciones en los campos de la salud, genética, economía, sociedad y marketing.

Los modelos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o descriptivos. Los modelos supervisados o predictivos se caracterizan por tener la variable de estudio ya determinada a priori, el objetivo en este caso es realizar predicciones en el conjunto de datos en base a las clasificaciones ya predeterminadas. Mientras que las técnicas no supervisadas o descriptivas son aquellas en las que no se dispone de clasificaciones a priori y más bien el objetivo es explorar el conjunto de datos para encontrar alguna estructura o forma de organizarlos.

Figura N° 2.1 Clasificación de Técnicas de Minería de Datos



Fuente: *Adaptado de la Figura 1 de Moreno, Quintales, García y Martin (2010)*

En la Figura N° 2.1 se muestran algunas de las técnicas de aprendizaje supervisado y no supervisado más utilizadas de la minería de datos. Dentro del segmento de técnicas supervisadas, encontramos dos bloques: uno de predicción y

otro de clasificación, que es donde se encuentran los modelos de árboles de clasificación, *random forest* y *adaboost,* utilizados en esta investigación.

2.2.2. Árbol de clasificación

Dentro de los modelos de aprendizaje supervisado más utilizados en la minería de datos, nos encontramos con los árboles de decisión, los cuales son principalmente utilizados para la exploración inicial y búsqueda de patrones en conjuntos de datos grandes. Pueden ser utilizados para problemas de clasificación y regresión.

El modelo utilizado para conocer los patrones de consumo de las TIC en los hogares del Perú es el modelo *Classification And Regression Tree* (CART), desarrollado por Leo Breiman en 1984.

El modelo CART consiste en la construcción de un árbol principal a partir de particiones binarias que se realizan de manera sucesiva en los datos mediante un criterio de particionamiento denominado medida de impureza, este criterio nos permite determinar la calidad de un nodo en la construcción del modelo de árbol de clasificación, y establecerá el grado de homogeneidad existente entre los grupos clasificados respecto a la variable dependiente. Luego en cada nodo construido la variable independiente (o predictora) que mejore más el criterio de particionamiento, es el que se usa para hacer la siguiente partición, los árboles que son creados con este criterio se dejan crecer ampliamente y después son podados para encontrar el tamaño óptimo del árbol (Pérez y Santín, 2007).

El algoritmo del modelo CART se podría resumir en tres grandes pasos (Timofeev, 2004):

1. Construcción del árbol principal o máximo
2. Poda del árbol.
3. Selección del árbol óptimo mediante un procedimiento de validación cruzada.

La medida de impureza o criterio de particionamiento será denotada por $i(l)$. Captura el grado en el que los casos dentro de un nodo están concentrados en una sola categoría. Si bien existen varias medidas de impureza, las dos más comunes son las siguientes (Breiman, 1984):

- El índice de entropía, que se define como:

$$i(l) = \sum_{i \neq j} p(j/l) lnp(i/l)$$

y cuyo objetivo es encontrar la partición que maximice $\triangle\, i(l)$:

$$\triangle\, i(l) = -\sum_{j=1}^{c} p(j/l) lnp(i/l)$$

donde $j = 1, \ldots, c$; es el número de clases de la variable respuesta (dependiente) y $p(j/l)$ la probabilidad de clasificación correcta para la clase j en el nodo l.

- El índice de Gini, el cual tiende a separar la categoría más grande en un grupo aparte, presenta la siguiente forma:

$$i(l) = \sum_{i \neq j} p(j/l)p(i/l)$$

Para luego encontrar la partición que maximice $\triangle\, i(l)$:

$$\triangle\, i = 1 - \sum_{j=1}^{C} (p_j)^2$$

donde $j = 1, \ldots, c$; es el número de clases de la variable respuesta categórica, $p(j/l)$ la probabilidad de clasificación correcta para la clase j en el nodo l y donde p_j es la frecuencia relativa de la clase c. Este índice es el utilizado en el modelo CART del software SPSS Versión 24, y también en los modelos *random forest* y *adaboost*, utilizados para la identificación de los factores asociados en el uso de las TIC, internet, telefonía fija y tv de paga de los hogares en el Perú, con el software R Studio versión 1.0.136.

El árbol que se obtiene inicialmente del modelo, por lo general se encuentra sobre ajustado, por lo cual se busca "podarlo" con el fin de encontrar el tamaño óptimo, cortando sucesivamente ramas o nodos terminales hasta encontrar el tamaño adecuado del árbol. Una forma de encontrar ese tamaño adecuado es la función de costo complejidad, cuyo proceso consiste en buscar una serie de árboles anidados de tamaños decrecientes (De'Ath, 2002), cada uno de los cuales es el mejor de todos los árboles de su tamaño, estos árboles pequeños son comparados para determinar el óptimo. La función de costo complejidad funciona bien si se deja crecer bastante el árbol, y se define para cada árbol de la siguiente manera:

$$Costo\ Complejidad = R(L) + \alpha * |L|$$

donde:

$R(L)$: Medida de riesgo de clasificación errónea del árbol o rama.

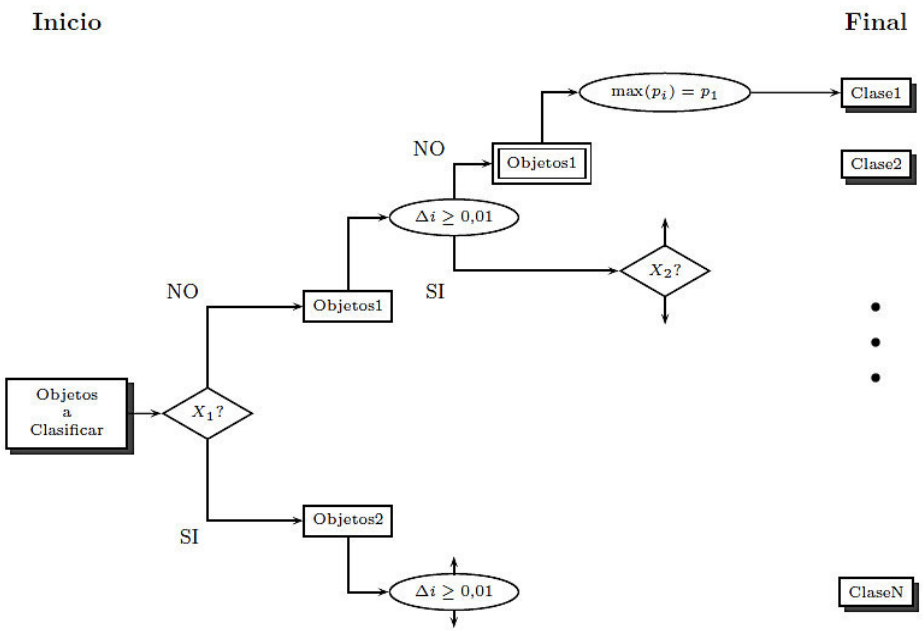
α: Coeficiente de penalidad.

$|L|$: Número de nodos terminales del árbol o rama.

Finalmente, para evaluar la bondad de la estructura de árbol cuando se generaliza para una mayor población, se utilizan principalmente la validación cruzada y validación mediante la matriz de error, exactitud, sensibilidad y especificidad. Éstos indicadores de validación, serán detallados con mayor detalle en la sección 2.2.5. de la presente investigación.

El algoritmo del modelo CART se resume en el esquema de la Figura N° 2.2:

Figura N° 2.2: Diagrama de flujo del algoritmo CART



Fuente: *Adaptado de la Figura 1.1 de Timofeev, R. (2004)*

2.2.3. Random forest

Modelo de minería de datos propuesto inicialmente por Kam Ho (1995) de Laboratorios Bell, y posteriormente desarrollado por Breiman (2001) quien presentó el modelo totalmente desarrollado.

Random forest surge como combinación de las técnicas de *Classification And Regression Tree* (CART) y *Booststrap Aggregating (Bagging)* para realizar la combinación de árboles predictores en la que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. *Bagging* también fue propuesto por Breiman (1994), basado en la combinación de clasificadores inestables como redes neuronales o árboles de clasificación (donde ligeros cambios en el conjunto de entrenamiento llevan a construir otro clasificador), la idea central de *bagging* es la de entrenar muchos clasificadores débiles independientes, para luego combinarlos todos en un clasificador fuerte, usando muestreo con reemplazamiento en el conjunto de datos.

El modelo *random forest* mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual, esta aleatorización puede introducirse en la partición del espacio (construcción del árbol), tanto en las etapas de entrenamiento, convalidación y de prueba.

En términos general el algoritmo del modelo *random forest* se desarrolla en los siguientes pasos (Montillo, 2009):

1. Para cada árbol b, donde b= 1 a B:

   (a) Se realiza la extracción de una muestra bootstrap $\mathbf{Z}^*$ de tamaño N a partir de los datos de entrenamiento. Al seleccionarse aleatoriamente con reemplazo, no todos los datos de conjunto general estarán en el conjunto de entrenamiento y no serán usados para crear el árbol, estos datos que no formen parte del conjunto de entrenamiento serán parte de un conjunto denominado *Out of bag data* (OOB *data*), con el cual se calculará el error de clasificación del modelo *random forest*.

   (b) Construir un árbol *random forest* T_b para los datos bootstrap, repitiendo de forma recursiva los siguientes pasos para cada nodo terminal del árbol, hasta que se alcanza el mínimo nodo de tamaño n_{min}.

      i. Seleccione m variables al azar de las p variables.
      ii. Elige la mejor variable / punto de división entre las m variables.
      iii. Dividir el nodo en dos nodos hijas.

2. Salida del conjunto de árboles $\{T_b\}_1^B$

   Para hacer una predicción en un nuevo punto x:

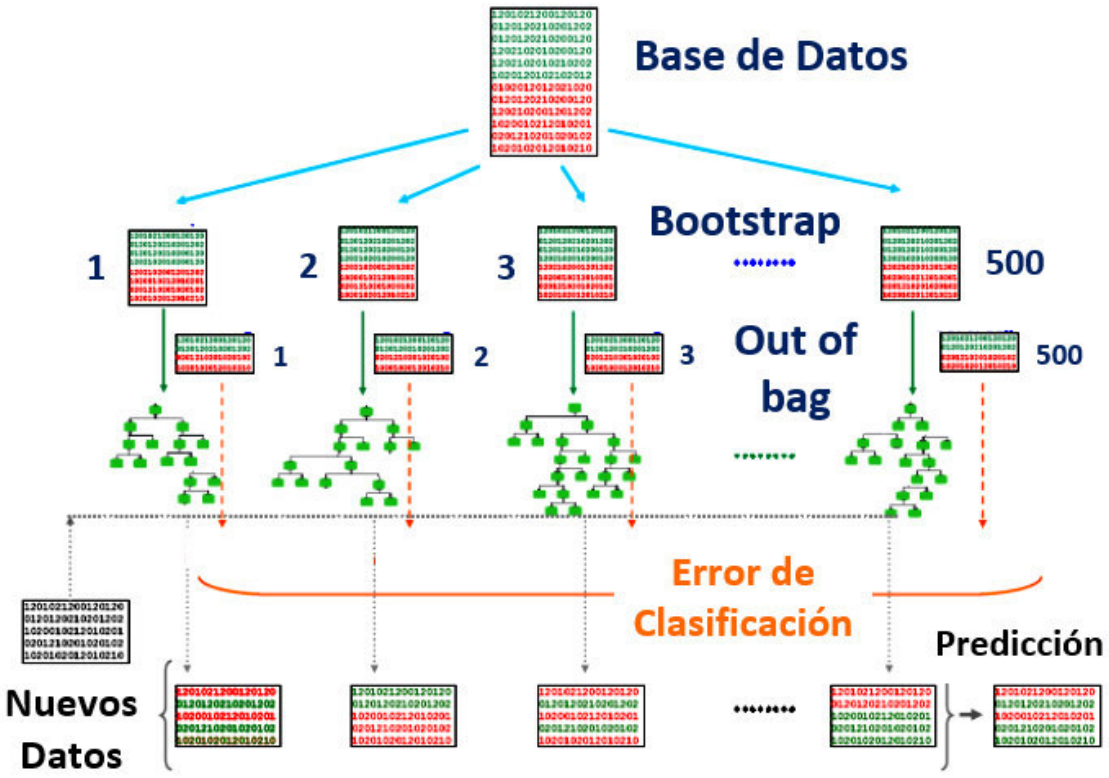
   $\hat{C}_b(x)$ es la predicción de la clase del b-ésimo árbol random forest. Entonces $\hat{C}_{rf}^B(x) = voto\ mayoritario\ \{\hat{C}_b(x)\}_1^B$. El voto mayoritario se

encarga de clasificar un nuevo dato, en base a la predicción realizada por la mayoría de árboles del modelo.

3. El error de clasificación del modelo se calcula a partir del conjunto de datos OOB. Se empieza a calcular el error de clasificación de cada iteración, a medida que se van agregando árboles al modelo *random forest*. Finalmente, el error del modelo se calculará como un promedio de los errores obtenidos en cada iteración.

El algoritmo del modelo *random forest* se resume en el esquema que se observa en la Figura N° 2.3:

Figura N° 2.3: Diagrama de flujo del algoritmo random forest



Fuente: *Adaptado de la Figura 1.2 de K. Hultstrom (2013)*

Dentro de las características más importantes del algoritmo del *modelo random forest* se podría mencionar las siguientes (Hultstrom, 2013):

- No se genera un único árbol, sino un gran número de ellos, éstos se construyen a partir de muchos conjuntos de datos similares generados mediante bootstrap (remuestreo con reposición) de la muestra original, así se consigue corregir el error de predicción debido a la selección específica del conjunto de datos y disponer para cada árbol de una muestra independiente para la estimación del error de clasificación, puesto que aproximadamente un tercio de la muestra original queda excluida de cada muestra generada por *bootstrap*.

- La aleatoriedad en este modelo es introducida para cada división de un nodo, es decir no se selecciona la mejor variable de entre todas, sino que se selecciona al azar un subconjunto de las p variables y se restringe la selección de la variable a este subconjunto. De esta forma se incluye una mayor variabilidad de árboles y se reduce la dependencia del resultado con las divisiones precedentes.

- *Random forest* establece rankings de importancia de las variables en la predicción de la variable respuesta.

Para identificar las variables de importancia, el modelo *random forest* tiene en consideración dos medidas (Williams, 2011):

- *Mean decrease accuracy* (MDA), basada en el aporte de la variable al error de clasificación (porcentaje o número de incorrectamente clasificados). El error de clasificación de cada árbol se calcula a partir de la parte de la muestra que ha quedado excluida de la submuestra utilizada en la construcción del árbol, generada por remuestreo; la diferencia, el número de correctamente clasificados (R_{OOB}). Para calcular la importancia de cada una de las variables que aparecen en un árbol, se permutan aleatoriamente los valores de esa variable, dejando intactos el resto de variables, y se vuelven a clasificar los mismos individuos según el mismo árbol, pero ahora con la variable permutada. La importancia en ese árbol se calcula como la diferencia entre el número de clasificaciones correctas antes de la permutación (R_{OOB}) y después de la permutación (R_{perm}) resultante. Finalmente se calcula la medida MDA, como la media de estas diferencias en todos los b-ésimos árboles, $b = 1, \dots, B$ en donde interviene la variable.

$$MDA = \frac{1}{B}\sum_{b=1}^{B}(R_{OOB} - R_{perm})$$

- *Mean decrease Gini* (MDG), calculada a partir del índice de Gini. Éste es el criterio que se utiliza para seleccionar la variable en cada partición en la construcción de los árboles y que constituye una disminución de esta medida. La importancia de una variable en un árbol se mide como la suma de los decrementos atribuidos a esa variable y la importancia final, como la media en todos los árboles.

$$MDG = 1 - \sum_{j=1}^{C} (p_j)^2 \quad (v)$$

donde $j = 1, \ldots, c$; es el número de clases de la variable respuesta categórica y p_j es la frecuencia relativa de la clase C en el modelo.

2.2.4. Adaboost

Adaboost es un modelo propuesto por Freund y Shapire en 1995, en el paper *"A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting"* con el cual ganaron el Premio Gödel en 2003. El nombre *Adaboost* viene de la mezcla entre *Adaptive* y *Boosting*, que son las dos características principales que ocupa el algoritmo.

Los modelos *boosting* buscan encontrar un clasificador general o "fuerte", en base a la combinación de una serie de clasificadores específicos o débiles (en general son reglas muy simples), además a diferencia de las técnicas *bagging*, que hace un remuestreo aleatorio de los datos de entrenamiento, *boosting* pondera las muestras para concentrar el aprendizaje en los casos más difíciles. Intuitivamente, se sabe que los casos más cercanos a la frontera de decisión son más difíciles de clasificar, por lo que recibirán pesos más altos. Se dice *Adaptive* (o Adaptivo), porque el clasificador "fuerte" se va adaptando y seleccionando los clasificadores débiles en función de los resultados que va obteniendo durante el entrenamiento. Este es un clasificador de tipo estadístico ya que usa características que se basan en la información cuantitativa de uno o varios elementos a estudiar (Williams, 2011).

A diferencia de *random fores*t, el proceso de formación *adaboost* selecciona sólo aquellas características conocidas para mejorar el poder predictivo del modelo, la reducción de dimensionalidad y potencialmente mejorar el tiempo de ejecución como características irrelevantes no necesitan ser calculado.

El algoritmo del modelo *adaboost* se desarrolla de la siguiente manera (Freund y Shapire, 1995):

- Determinar el número de iteraciones a realizar e inicializar la distribución de pesos otorgados a cada uno de datos $D_b(i) = 1/N$.
- Para b que es cada árbol individual y donde $b = 1, \dots, B$.

1. Entrenar al clasificador débil $h_b(x)$ usando la distribución D_b.
2. Elegir un peso (valor de confianza) $\alpha_b \in R$.
3. Actualizar la distribución sobre el conjunto de entrenamiento:

$$D_{b+1}(i) = \frac{D_b(i) e^{-\alpha_b y_i h_t(x_i)}}{Z_b}$$

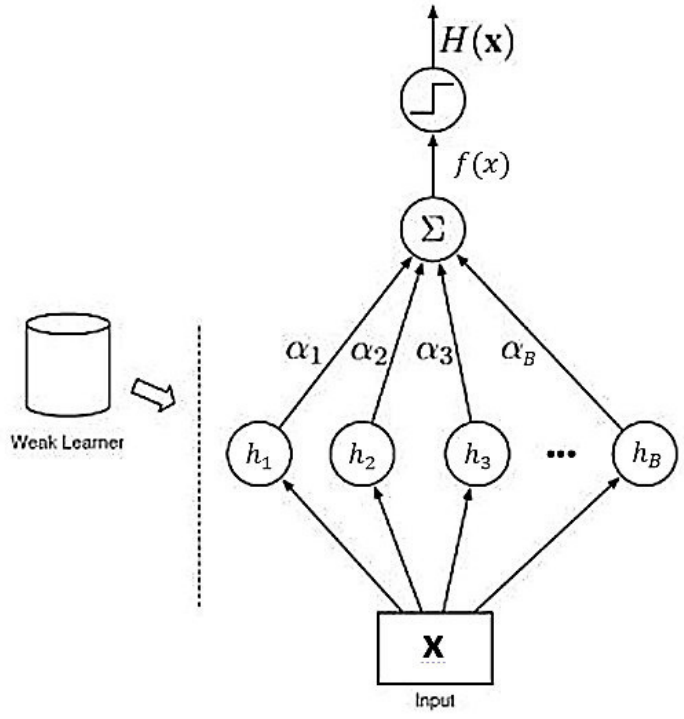
- Set $H(x) = sign\big(f(x)\big) = sign(\sum_{i=1}^{B} \alpha_b h_b(x))$

donde $h_b(x)$ es un clasificador débil, α_t un factor que combina linealmente los clasificadores débiles. La respuesta final del clasificador se toma como $sign\big(f(x)\big)$, donde un valor mayor que 0, representa una detección positiva del objeto y un valor

menor o igual a 0, representa que no se encuentra el objeto en la ventana analizada. En síntesis se puede mencionar que el funcionamiento del modelo *adaboost* emplea un factor $\alpha_b = \frac{1}{2}\ln\left(\frac{1-\varepsilon_b}{\varepsilon_b}\right)$, donde ε_b es el error asociado a h_b, dicho factor es utilizado por el modelo para combinar los clasificadores $h_b(x)$ y determinar las distribuciones D_b. Además α_b, definido en función del error ε_b, mide la relevancia de h_b en el clasificador final $H(x)$, es importante resaltar que α_b es mayor cuanto menor es ε_b, dando así mayor importancia a aquellas hipótesis que cometen menos errores de clasificación. *Adaboost* escala el peso $D_b(i)$ mediante la expresión $e^{-\alpha_b y_i h_b(x_i)}$, aumentando el peso los casos mal clasificados.

El algoritmo del modelo *Adaboost* se resume en el esquema que se observa en la Figura N° 2.4:

Figura N° 2.4 Diagrama de flujo del algoritmo del modelo adaboost



Fuente: *Adaptado de la Figura 1 de Bailly y Milgram (2009)*

Para determinar la importancia de las variables, al igual que el modelo *random forest*, *adaboost* se basa en la ganancia del índice de Gini para el cálculo de importancia de cada variable p obtenida en cada árbol b, tomando en consideración el peso del árbol (Alfaro, Gamez & García, 2013). Esta medida de importancia está definida de la siguiente manera:

$$Imp_p^2(b) = \sum_{l=1}^{L} \hat{d}_l \cdot 1\{dividir\ el\ nodo\ l\ sobre\ la\ variable\ p\}$$

Donde L es el número de nodos internos, y $\hat{d}_l$ es la mejora en el error de clasificación en el error de entrenamiento al hacer la l-ésima división.

Finalmente, la importancia de variables con el modelo *adaboost* se calcula como un promedio simple de la importancia al cuadrado de todos los árboles individuales. Esta importancia se presenta de la siguiente forma:

$$Imp_p(T_b) = \sqrt{\frac{1}{B}\sum_{b=1}^{B} Imp_p^2(b)}$$

Este promedio estabiliza la importancia de variables calculada por cada árbol de manera individual, lo que genera que esta medida tienda a ser mucho más precisa que otras medidas de importancia.

2.2.5. Evaluación de los modelos

Cuando se utilizan modelos de minería de datos en una investigación, la muestra en estudio es segmentada en 3 partes con el fin de garantizar la calidad en el proceso del modelo:

- Muestra de Entrenamiento (*Training*): son los datos con los que se entrenan los modelos.
- Muestra de Prueba: selecciona el mejor de los modelos entrenados.
- Muestra de Validación: Entrega el error real cometido con el modelo seleccionado.

Durante el proceso de selección del mejor modelo, los modelos se ajustan a los datos de entrenamiento y el error de predicción para dichos modelos es obtenido mediante el uso de los datos de prueba. Finalmente, una vez que termina el proceso y se tiene seleccionado el modelo, se utilizan los datos de validación para evaluar la manera en que el modelo seleccionado se generaliza para los datos que no jugaron ningún papel en la selección del mismo.

Para realizar la evaluación de los resultados obtenidos por los modelos *adaboost* y *random forest*, se utilizaron los siguientes criterios:

2.2.5.1. Matriz de error: o también conocida como matriz de confusión, muestra los resultados reales de clasificación de la muestra contra los resultados de predicción clasificados por el modelo, la idea central de

validación mediante esta técnica es encontrar estructuras de clasificación similares, tanto para los datos utilizados en la fase de entrenamiento, prueba y validación. De ser las estructuras de clasificación similares, se considera que el modelo utilizado es eficiente para la clasificación. El uso de esta técnica de validación es adecuado cuando la variable de destino (variable dependiente) es categórica. (Williams, 2011). A continuación, se muestra la distribución de resultados según la matriz de error:

		Predicho	
		Sí	**No**
Real	**Sí**	VP	FN
	No	FP	VN

Donde:

VP: Verdaderos Positivos, se refiere a aquellas observaciones predichas positivas (Sí) clasificadas de manera correcta.

FN: Falsos Negativos, se refiere a aquellas observaciones predichas positivas (Sí) clasificadas de manera incorrecta.

VN: Verdaderos Negativos, se refiere a aquellas observaciones predichas negativas (No) clasificadas de manera correcta.

FP: Falsos Positivos, se refiere a aquellas observaciones predichas negativas (No) clasificadas de manera incorrecta.

2.2.5.2. Exactitud: Medida de calidad que permite encontrar el porcentaje total de aciertos de predicción de la clasificación del modelo en relación con el total de observaciones reales. La tasa de error de clasificación total del modelo resulta de la diferencia del porcentaje total de casos

clasificados y el valor obtenido con la exactitud. La exactitud presenta la siguiente fórmula:

$$Exactitud = \frac{Total\ de\ observaciones\ correctamente\ predichas}{Total\ de\ observaciones}$$

$$= \frac{VP + FP}{VP + VN + FP + FN}$$

2.2.6. Comparación de Modelos

2.2.6.1. Sensibilidad: Medida de calidad que permite encontrar la proporción de observaciones positivas, correctamente clasificadas por el modelo frente al total de observaciones positivas reales. La tasa de error de clasificación de las observaciones positivas del modelo resulta de la diferencia del porcentaje total de casos positivos y el valor obtenido con la sensibilidad. La sensibilidad presenta la siguiente fórmula:

$$Sensibilidad = \frac{Total\ de\ observaciones\ positivas\ correctamente\ predichas}{Total\ de\ observaciones\ positivas}$$

$$= \frac{VP}{VP + FN}$$

2.2.6.2. Curva ROC: El Análisis ROC (*Receiver operating characteristics*) es una metodología desarrollada para evaluar la capacidad de un modelo para clasificar de manera correcta. Un valor de 1 significa que el método es perfecto; un valor de 0.5 indica que el método no es útil, y valores intermedios miden la capacidad del método para discriminar. Una de las

principales ventajas de usar las curvas ROC es que además de entregarnos un valor de decisión de manera automática, nos muestra una representación gráfica de fácil interpretación y rápido entendimiento visual (Aler, 2014).

III. METODOLOGÍA

3.1. Tipo y diseño de investigación

La investigación presentada en esta tesis es de tipo aplicada, puesto que el propósito es dar solución a situaciones o problemas concretos e identificables (Bunge, 1971); cuantitativa, ya que se obtuvieron resultados a partir del uso de herramientas estadísticas e informáticas; y descriptiva, puesto que con esta investigación llegaremos a conocer características (factores) asociados a los patrones de uso de las TIC en los hogares del Perú. El diseño de la investigación es no experimental, debido a que no se realizaron manipulaciones de variables, sino que se trabajó a partir de factores ya existentes, y de corte transversal, porque los datos fueron recolectados de manera única en el tiempo.

3.2. Población, unidad de análisis

La población con la que se trabajó esta investigación son los hogares del Perú, ubicados en todas las áreas urbanas y rurales durante el año 2014. La unidad de análisis es cada hogar ubicado a lo largo del territorio nacional.

3.3. Tamaño y selección de la muestra

Se trabajó con una muestra de 14841 hogares de todo el Perú, tanto de áreas urbanas como rurales, la muestra seleccionada fue de tipo probabilística, multietápica, estratificada, por conglomerados estratificados implícitamente por nivel socio económico y de selección sistemática, a un nivel de confianza del 95%.

- Marco Muestral: Censo de Población y Vivienda 2007.
- Unidad de Muestreo:
    - La Unidad Primaria de Muestreo (UPM) es el centro poblado.
    - La Unidad Secundaria de Muestreo (USM) en el caso urbano son los conglomerados (agrupación de viviendas contiguas que generalmente forman "manzanas" completas). En el caso rural, los centros poblados están constituido por viviendas, contiguas o dispersas.
    - La Unidad Terciaria de Muestreo (UTM) solo existe en el caso urbano y es la vivienda particular.

3.4. Variables

Se utilizaron un total de 16 variables, 12 independientes relacionadas a características del jefe del hogar y la vivienda y 4 variables dependientes relacionadas al uso de las TIC, internet, tv de paga y telefonía fija.

Variables dependientes (o de destino):

- Uso de las TIC

- Uso del Internet
- Uso de la tv de paga
- Uso de la telefonía fija.

Variables independientes (o predictoras):

- Área sociodemográfica
- Nivel socioeconómico del hogar
- Nivel de pobreza del hogar
- Condición de la vivienda
- Número de miembros del hogar
- Número de habitaciones de la vivienda
- Departamento
- Edad del jefe del hogar
- Tiempo de ocupación en la vivienda
- Nivel educativo del jefe del hogar
- Lengua materna del jefe del hogar
- Sexo del jefe del hogar

3.5. Fuente de información

La fuente de datos utilizada para esta investigación es de tipo secundaria y corresponde a la ENCUESTA RESIDENCIAL DE SERVICIOS DE TELECOMUNICACIONES ERESTEL 2014, realizada por el Organismo Supervisor de Inversión Privada en Telecomunicaciones (OSIPTEL). Esta encuesta tiene como objetivo principal obtener información sobre demanda y patrones de uso de los

servicios de telecomunicaciones en el Perú.

3.6. Softwares utilizados

Los softwares utilizados en esta tesis son el software libre R Studio versión 1.0.136., se utilizó su paquete Rattle, para desarrollar los análisis mediante los modelos de minería de datos *random forest* y *adaboost*, es importante resaltar que el paquete Rattle presenta una interfaz, que permite al usuario aplicar modelamiento de datos y análisis estadísticos, sin la necesidad de realizar programación, usando el mismo entorno del software R Studio. Para el desarrollo del modelo de árbol de clasificación se utilizó el software IBM SPSS Statistics versión 24, con licencia, adquirido por la Universidad Nacional Mayor de San Marcos.

IV. RESULTADOS

4.1. Análisis Descriptivo:

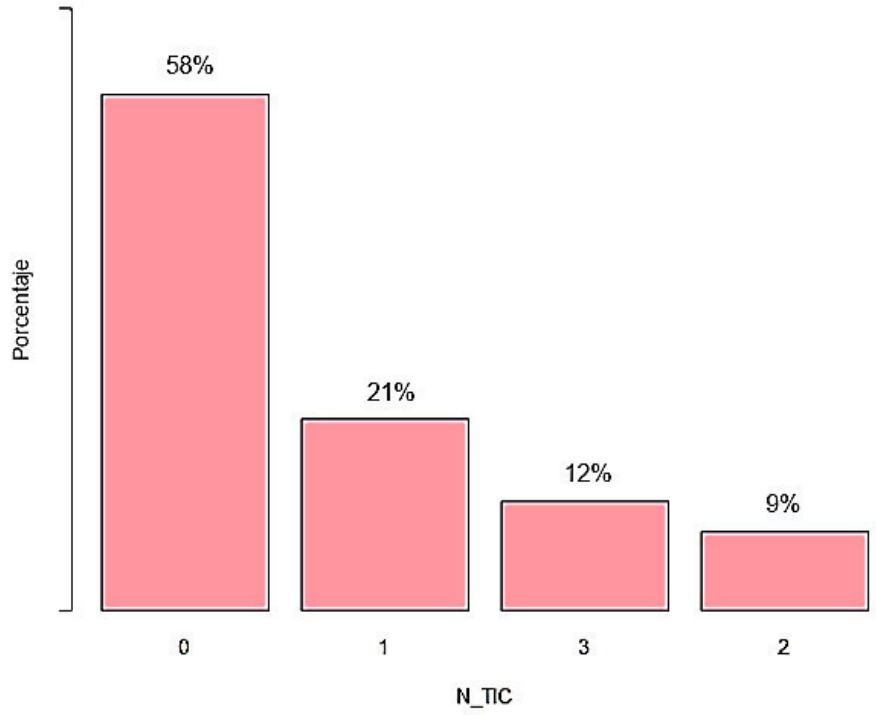
En relación con los hogares estudiados se puede indicar que el 75,7% de ellos, pertenecen a áreas urbanas dispersas a lo largo del territorio nacional, además que en respecto a los niveles de pobreza, el 25,6% de los hogares peruanos son considerados económicamente pobres, siendo el nivel socioeconómico D, en el que se encuentran con mayor proporción, con un 38,4%.

Respecto a la ubicación geográfica, es importante resaltar que el 18,8% de los hogares que fueron estudiados se encuentran ubicados en el departamento de Lima, con un 13,7% ubicado en los diferentes distritos de Lima Metropolitana y un 5,1% en las otras provincias de Lima. Otro departamento con proporción importante de hogares considerados para el estudio son Tumbes con un 6,6% y Puno con 6,2%.

Para el caso de los jefes de hogar, es importante resaltar que el 23% de ellos son mujeres, además que sólo el 26,7% de los jefes de hogar cuentan con estudios superiores, ya sean de tipo universitario o técnico. Respecto a la edad de los jefes de hogar entrevistados, se puede comentar que la mayoría pertenecen a la generación X (Entre 35 a 49 años) y Boomers (Entre 50 a 64 años) con una proporción de 34,1% y 29% respectivamente. Respecto a la lengua materna de los jefes de hogar, en el 85% de ellos es el castellano, mientras que en un 13% lo es el quechua.

En relación con el uso de las TIC en los hogares del Perú, se puede comentar de la Figura N° 4.1, que el 42% de los hogares peruanos usa por lo menos uno de los servicios TIC empaquetados en estudio (Teléfono fijo, Internet, TV de Paga), mientras que en solo un 12% de los hogares se hace uso de los 3 servicios TIC mencionados.

Figura N° 4.1: Distribución de hogares del Perú por N° de TIC utilizadas, 2014

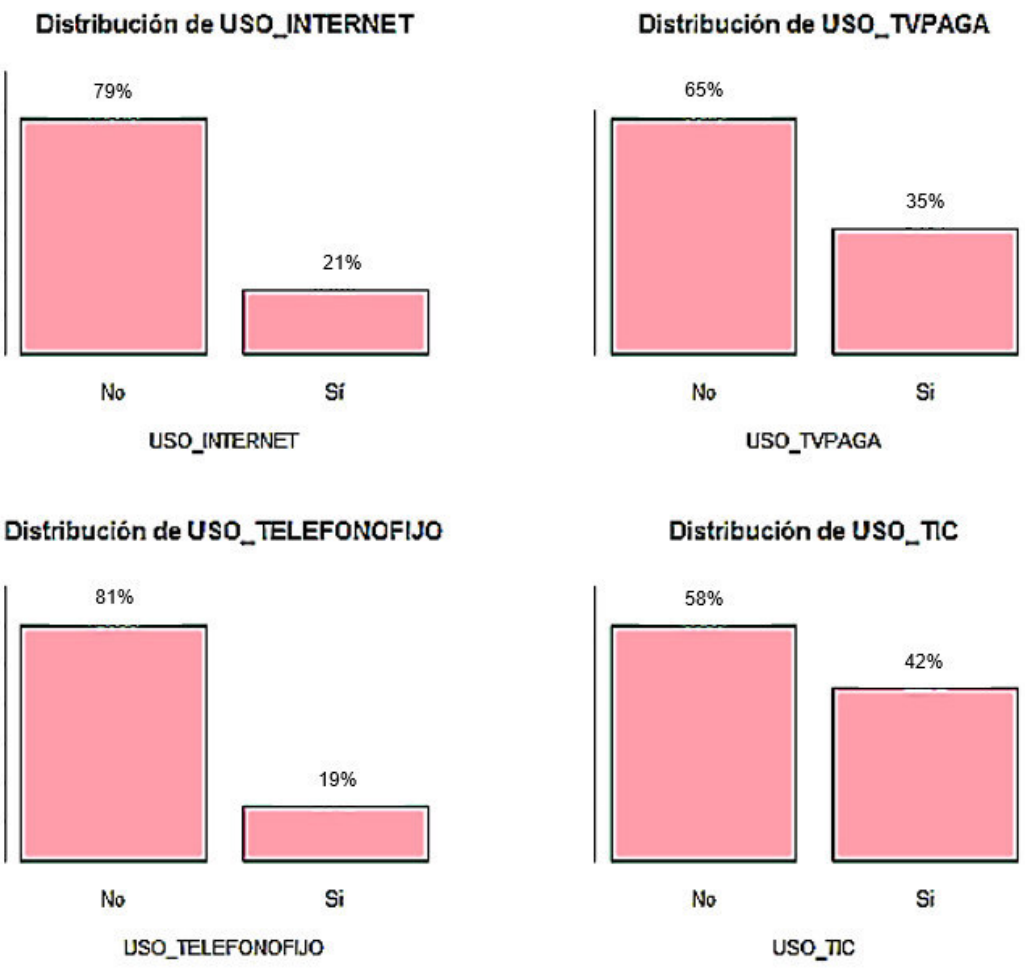


***Fuente**: Elaboración Propia*

En relación con los usos de los TIC en forma individual, de la Figura N° 4.2, se observa que el 19,1% de los hogares peruanos sí hace uso de la telefonía fija, el 21,3% sí hace uso del Internet fijo ya sea desde su computadora de mesa o portátil y el 34,8% de los hogares en Perú, sí hace uso de la TV de Paga. Además se puede apreciar que para todos los casos de usos de tecnologías estudiados de forma individual, es mayor el número de hogares que cuentan con estas tecnologías, lo cual

se puede explicar con la gran demanda que vienen teniendo los celulares inteligentes (smartphones) a lo largo de los últimos años, y que puede estar generando una nueva tendencia de migración de los usos de servicios de tecnologías fijos por el uso de las nuevas herramientas tecnológicas móviles, que hoy en día permiten realizar las diferentes actividades que se realizan con las TIC "fijas" del hogar, con la gran ventaja que pueden ser realizadas no sólo dentro del hogar, sino desde el lugar en el que te encuentres.

Figura N° 4.2: Distribución de hogares del Perú según uso de las TIC, 2014



Fuente: *Elaboración Propia*

4.2. Modelos de Minería de Datos

Se utilizaron los modelos *random forest* y *adaboost* para identificar los factores asociados de los hogares del Perú tanto para el uso de las TIC, internet, tv de paga y telefonía fija, consideradas como variables de destino (variables dependientes), con el software libre de programación estadística R Studio versión 1.0.136.

Para cada uno de los modelos utilizados, se distribuyó la muestra en estudio en 3 partes: 70% (10389 hogares) para la etapa de construcción del modelo (muestra de entrenamiento), 15% (2226 hogares) para la etapa de prueba y otro 15% (2226 hogares) para la etapa de validación. Durante la etapa de entrenamiento de cada uno de los modelos, se crearon 500 árboles, al producirse la división en cada nudo del árbol se usaron los 12 predictores (variables independientes) mostrados en la matriz de operacionalización (Anexo N° 2), relacionadas a las características del hogar y al jefe del hogar como variables de entrada. Además de las validaciones del modelo con el análisis de resultados obtenidos en las etapas de prueba y validación, se utilizó la sensibilidad y el análisis mediante la curva ROC para la elección del mejor modelo.

4.2.1. Uso de las TIC

Con el modelo *random forest* se obtuvo una exactitud 76%, mientras que con el modelo *adaboost* se clasificó correctamente al 79% de los hogares del Perú. Además, se observa mayores resultados de sensibilidad y ROC con el modelo *adaboost*, tal como se observa en la Tabla N° 4.1.

Tabla N° 4.1: Comparación de modelos Random Forest Vs. Adaboost

Variable de destino: Uso de las TIC

Indicadores	Random forest			Adaboost		
	Entrenamiento	Prueba	Validación	Entrenamiento	Prueba	Validación
Tamaño de la Muestra	10389 (70%)	2226 (15%)	2226 (15%)	10389 (70%)	2226 (15%)	2226 (15%)
Sensibilidad del modelo	70%	69%	71%	75%	74%	74%
ROC	83%			88%		

Fuente: *Elaboración Propia*

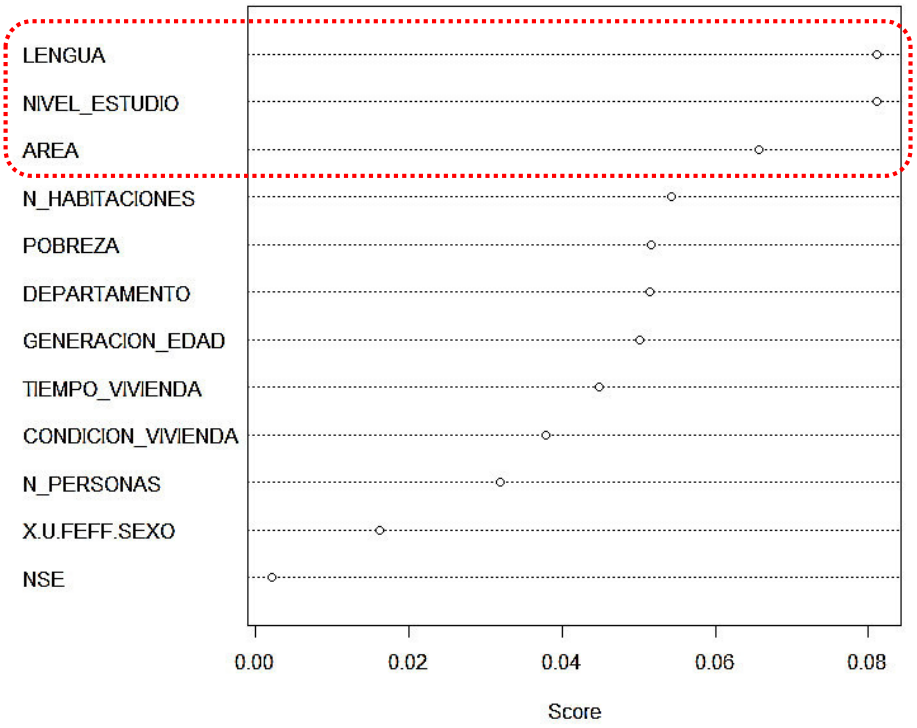
En general, con ambos modelos se obtuvieron buenos resultados de clasificación, pero para la identificación de los factores asociados al uso de las TIC, se utilizó los resultados obtenidos mediante el modelo *adaboost,* que es con el que se obtiene mejores resultados en la sensibilidad y ROC en la clasificación de hogares del Perú.

En la Figura N° 4.3 se presentan las valoraciones de las variables por "importancia", es decir el nivel de influencia de cada una de las variables independientes sobre la variable dependiente (Uso de las TIC) en el modelo *adaboost*. La forma de interpretación de esta gráfica se da respecto a las valoraciones obtenidas por las variables, es decir, cuanto mayor sea la magnitud de la valoración, "más importante" será la variable correspondiente. Si bien no existe un criterio definido que indique que cantidad de variables deben ser elegidas cómo las más importantes. Landgraf (2012) indica que si bien las variables más importantes en el modelo, obtenidas mediante los modelos *random forest* y *adaboost* carecen de alguna interpretabilidad definida, dependerá del investigador o del área de investigación en el que nos encontremos aplicando los modelos de minería de datos

para tomar una decisión al respecto. Es por eso por lo que, para esta investigación, se consideraron como los factores asociados al uso de las TIC en los hogares del Perú en el 2014, las variables independientes con una contribución superior al 10%.

Figura N° 4.3: Importancia de variables en el modelo Adaboost

Variable de destino: Uso de las TIC



Fuente: *Elaboración Propia*

De la Figura N° 4.3 y Tabla N° 4.2, podemos observar que son la lengua materna, el nivel de estudio del jefe del hogar y el área sociodemográfica donde se encuentra ubicada la vivienda los más importantes factores asociados al uso de las TIC en los hogares del Perú en el 2014. Además de la Tabla 4.26 se observa que sólo considerando las 3 variables seleccionadas como de "mayor importancia", se tiene una contribución del 39% en el modelo *adaboost* desarrollado.

Tabla N° 4.2: Importancias y contribuciones de las variables independientes con el modelo adaboost (Variable de destino: Uso de las TIC)

VARIABLE	Importancia	%
Lengua materna	0,082	13,9%
Nivel de estudio	0,082	13,9%
Área sociodemográfica	0,066	11,2%
N.º de habitaciones	0,054	9,2%
Nivel de pobreza	0,052	8,8%
Departamento	0,052	8,8%
Edad del jefe del hogar	0,050	8,5%
Tiempo en la vivienda	0,045	7,6%
Condición de la vivienda	0,038	6,5%
N.º de personas	0,032	5,4%
Sexo	0,016	2,7%
Nivel socioeconómico	0,020	3,4%

Fuente: *Elaboración Propia*

4.2.1.1. Patrones de consumo mediante árboles de clasificación

Si bien se identificó que son la lengua materna, el nivel de estudio del jefe del hogar, el área sociodemográfica donde se encuentra ubicada la vivienda y el número de habitaciones que posee la misma, los más importantes factores asociados al uso de las TIC en los hogares del Perú en el 2014, también es necesario conocer cuáles son las características más importantes de estos factores para que se determine usar las TIC en el hogar, para ello se utilizó el modelo de árbol de clasificación CART.

Para la construcción de este modelo se consideró como variable dependiente, el uso de las TIC, y como variables independientes a los 4 factores más importantes asociados al uso de las TIC, identificados con el modelo *adaboost*.

De la Tabla N° 4.3 se observa que, durante la etapa de validación, se obtuvo una exactitud del 71% con el modelo de árbol de clasificación CART. Además, es importante resaltar que se obtuvieron valores de sensibilidad del 81% y ROC del 78%.

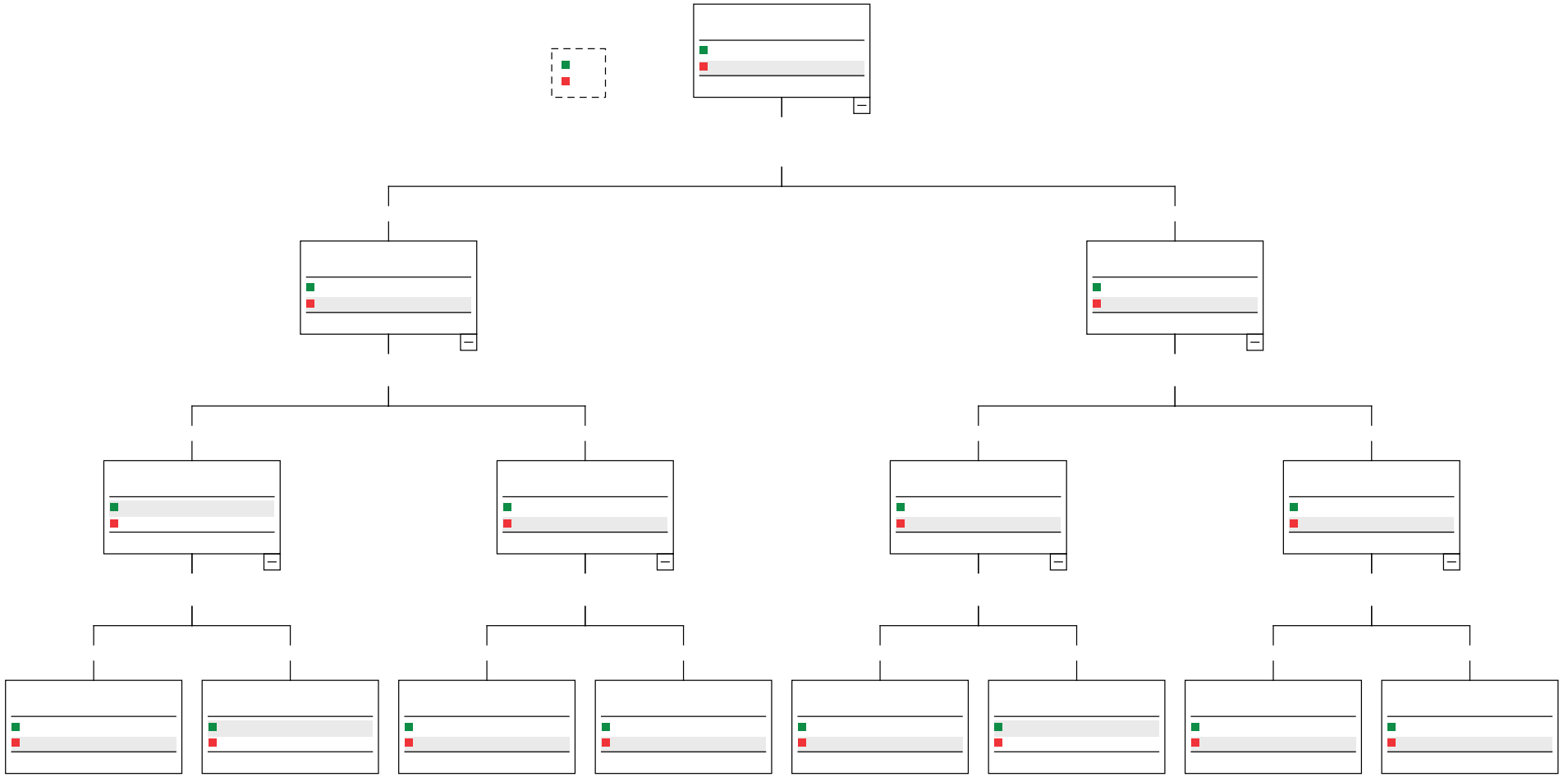
Tabla N° 4.3: Validación del modelo de árbol de clasificación CART.

Variable de destino: Uso de las TIC

Etapa	Real	Predicho		Exactitud	Sensibilidad	ROC
		Si	No			
Entrenamiento	Si	3438	893	68,91%	79,40%	75,36%
	No	2324	3691			
Validación	Si	1566	379	70,55%	80,51%	77, 96%
	No	945	1606			

Fuente: *Elaboración Propia*

Figura N° 4.4: Árbol de clasificación para identificar los patrones de consumo del uso de las TIC en los hogares del Perú, 2014.



Fuente: *Elaboración Propia*

En la Figura N° 4.4 observamos el diagrama obtenido con la construcción del modelo de árbol de clasificación, la interpretación que se muestra a continuación se realizó con los resultados obtenidos del modelo a partir de la muestra de validación:

- Del nodo 0 del árbol, encontramos que el 42% de los hogares del Perú sí hacen uso de las TIC.

- Del nodo 1 y 2, se observa que son los hogares del Perú, con jefes de familia que tienen como lengua materna el castellano, los que más hacen uso de las TIC. En el 46% de los hogares con jefes de hogar que tienen como lengua materna el castellano, se hace uso de las TIC; situación distinta a la que se observa en los hogares, con jefes de hogar cuya principal lengua es la quechua, aymara u otra nativa, donde solo el 14% hace uso de las TIC.

- Del nodo 3 y 4, enfocándonos sólo en los hogares con jefes de hogar que tienen como lengua materna el castellano, observamos que son los hogares que habitan en áreas urbanas, donde se hace uso de las TIC en mayor proporción, con un 53%, mientras que en los hogares que se encuentran en áreas rurales, la proporción es de sólo el 19%.

- Del nodo 7 y 8, enfocándonos sólo en los hogares con jefes de hogar que tienen como lengua materna el castellano y que habitan en áreas urbanas, observamos que son los hogares del Perú, con jefes de familia con niveles educativos superior (técnica o universitaria), los que más hacen uso de las TIC. El 71% de los hogares con jefes de hogar con nivel educativo superior, hacen uso de las TIC, situación contraria con lo que se encuentra en los

hogares con jefes de familia con menor nivel educativo, en donde sólo el 43% hace uso de las TIC.

Como interpretación general de los resultados obtenidos con el modelo de árbol de clasificación CART para identificar los patrones de consumo de los TIC en los hogares del Perú, se puede decir que son aquellos hogares con jefes de familia que tienen como lengua materna el castellano, niveles de educación superior (ya sea técnica o universitaria), además de encontrarse ubicadas en áreas urbanas del país, son los que más hacen uso de las TIC.

4.2.2. Uso de internet

Con el modelo *random forest* se obtuvo una exactitud del 83%, mientras que con el modelo *adaboost* se clasificó correctamente 86% de los hogares del Perú según el uso de internet, tal como se observa en la Tabla N° 4.4.

Tabla N° 4.4: Comparación de modelos Random Forest Vs. Adaboost

Variable de destino: Uso de Internet

Indicadores	Random forest			Adaboost		
	Entrenamiento	Prueba	Validación	Entrenamiento	Prueba	Validación
Tamaño de la Muestra	10389 (70%)	2226 (15%)	2226 (15%)	10389 (70%)	2226 (15%)	2226 (15%)
Sensibilidad del modelo	61%	60%	60%	64%	63%	62%
ROC (Validación)	86%			90%		

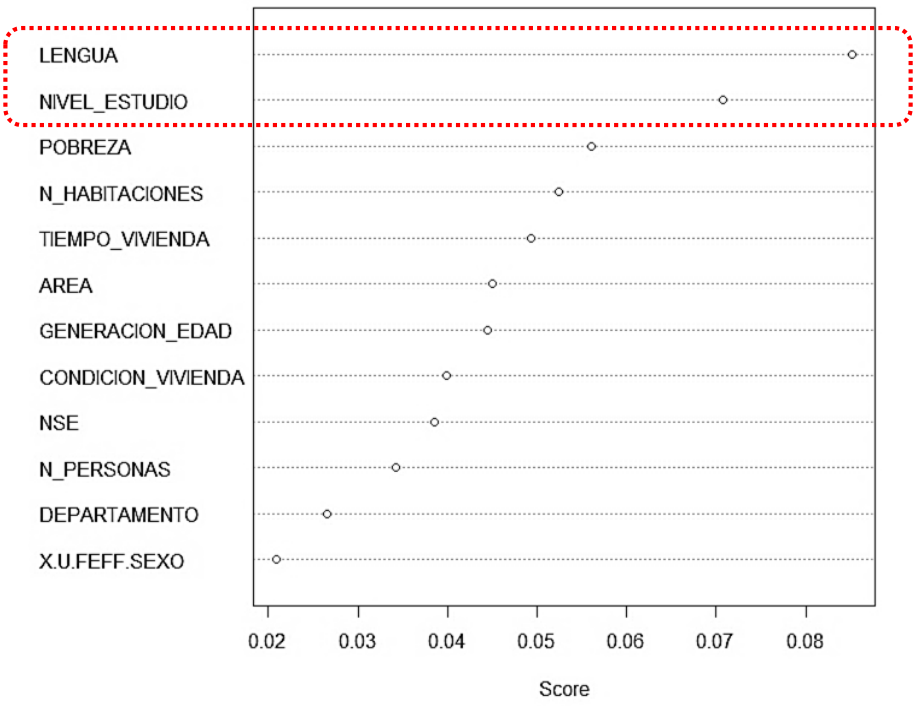
Fuente: *Elaboración Propia*

Con ambos modelos se obtuvieron buenos niveles de clasificación, pero para la identificación de los factores asociados al uso de internet, se utilizaron los

resultados obtenidos mediante el modelo *adaboost*, que es el que presenta mayores resultados de sensibilidad (62% en la validación) en la clasificación de hogares del Perú, además de obtener un valor de 90% con la curva ROC para este modelo, superior al obtenido con *random forest*.

Figura N° 4.5: Importancia de variables en el modelo Adaboost

Variable de destino: Uso de internet



Fuente: *Elaboración Propia*

De la Figura N° 4.5 y Tabla N° 4.5, podemos observar que son la lengua materna y el nivel de estudio del jefe del hogar, los más importantes factores asociados al uso de internet en los hogares del Perú en el 2014. Considerando sólo las dos variables clasificadas como "muy importantes", se obtiene una contribución en el modelo *adaboost* de alrededor 28,1%.

Tabla N° 4.5: Importancias y contribuciones de las variables independientes con el modelo adaboost (Variable de destino: Uso de Internet)

VARIABLE	Importancia	%
Lengua materna	0,086	15,5%
Nivel de estudio	0,071	12,6%
Nivel de pobreza	0,056	9,6%
N.º de habitaciones	0,053	9,4%
Tiempo en la vivienda	0,049	8,7%
Área sociodemográfica	0,045	8,0%
Edad del jefe del hogar	0,045	8,0%
Condición de la vivienda	0,04	7,1%
Nivel socioeconómico	0,038	6,7%
N.º de personas	0,034	6,0%
Departamento	0,027	4,8%
Sexo	0,021	3,7%

Fuente: *Elaboración Propia*

4.2.3. Uso de tv de paga

Se obtuvo con el modelo *adaboost*, 78% de exactitud en la clasificación de los hogares del Perú que hacen uso de la tv de paga.

Tabla N° 4.6: Comparación de modelos Random Forest Vs. Adaboost Variable de destino: Uso de tv de paga

Indicadores	Random forest			Adaboost		
	Entrenamiento	Prueba	Validación	Entrenamiento	Prueba	Validación
Tamaño de la Muestra	10389 (70%)	2226 (15%)	2226 (15%)	10389 (70%)	2226 (15%)	2226 (15%)
Sensibilidad del modelo	62%	62%	62%	64%	64%	65%
ROC (Validación)	81%			84%		

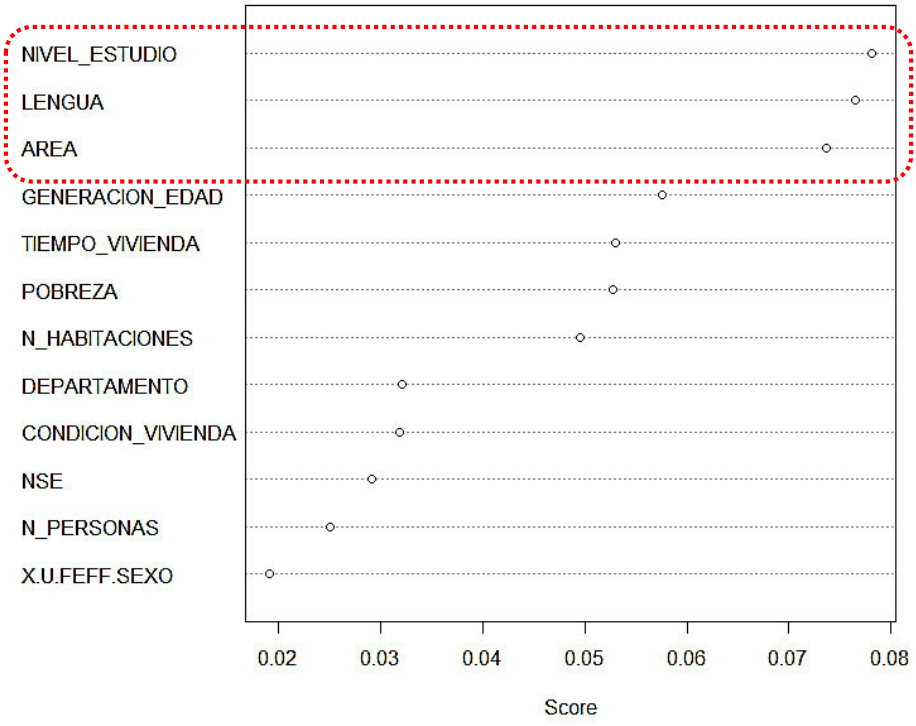
Fuente: *Elaboración Propia*

Tal como se observa en la Tabla 4.6, con el modelo *adaboost* se obtuvieron mayores valores de sensibilidad (65% en la validación) y de curva ROC (84%), por lo que es mediante este modelo que identificaremos los factores asociados al uso de tv de paga en los hogares del Perú.

De la Figura N° 4.6 y Tabla N° 4.6, podemos observar que son el nivel de estudio y la lengua materna del jefe del hogar, y el área sociodemográfica donde se encuentra ubicada la vivienda, los factores más importantes en el uso de tv de paga de los hogares del Perú.

Figura N° 4.6: Importancia de variables en el modelo Adaboost

Variable de destino: Uso de tv de paga



Fuente: *Elaboración Propia*

Además de la Tabla 4.7 se observa que sólo considerando las 3 variables clasificadas como "muy importantes" se tiene una contribución de 39,5% en el modelo *adaboost* desarrollado.

Tabla N° 4.7: Importancias y contribuciones de las variables independientes con el modelo adaboost (Variable de destino: Uso de tv de paga)

VARIABLE	Importancia	%
Nivel de estudio	0,078	13,6%
Lengua materna	0,076	13,2%
Área sociodemográfica	0,074	12,8%
Edad del jefe del hogar	0,057	9,8%
Tiempo en la vivienda	0,053	9,2%
Nivel de pobreza	0,053	9,2%
N.º de habitaciones	0,049	8,5%
Departamento	0,032	5,6%
Condición de la vivienda	0,032	5,6%
Nivel socioeconómico	0,029	5,0%
N.º de personas	0,025	4,3%
Sexo	0,018	3,1%

Fuente: *Elaboración Propia*

4.2.4. Uso de telefonía fija

Con el modelo *random forest* se obtuvo una exactitud del 85%, mientras que con el modelo *adaboost* la exactitud fue del 87%.

Tabla N° 4.8: Comparación de resultados Random Forest Vs. Adaboost Variable de destino: Uso de telefonía fija

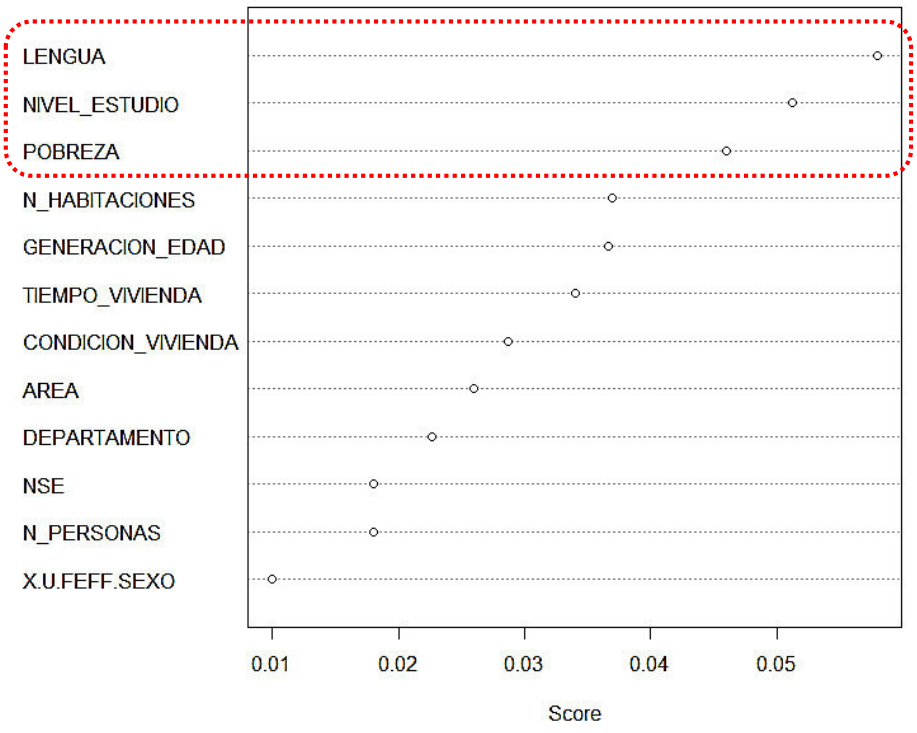
Indicadores	Random forest			Adaboost		
	Entrenamiento	Prueba	Validación	Entrenamiento	Prueba	Validación
Tamaño de la Muestra	10389 (70%)	2226 (15%)	2226 (15%)	10389 (70%)	2226 (15%)	2226 (15%)
Sensibilidad del modelo	60%	61%	61%	62%	62%	63%
ROC (Validación)	89%			91%		

Fuente: *Elaboración Propia*

De la Tabla 4.8, se observa que con el modelo *adaboost* se obtuvieron mayores valores de sensibilidad (63% en la validación) y de curva ROC (91%), por lo que es mediante este modelo que identificaremos los factores asociados al uso de la telefonía fija en los hogares del Perú. De la Figura N° 4.8 y Tabla N° 4.9, podemos observar que estos factores son la lengua materna, el nivel de estudio del jefe del hogar, y el nivel de pobreza del hogar.

Figura N° 4.8: Importancia de variables en el modelo Adaboost

Variable de destino: Uso de telefonía fija.



Fuente: *Elaboración Propia*

Además de la Tabla 4.9 se observa que, sólo considerando las 3 variables de mayor importancia, se tiene una contribución de 40,2% en el modelo *adaboost* desarrollado.

Tabla N° 4.9: Importancias y contribuciones de las variables independientes con el modelo adaboost (Variable de destino: Uso de telefonía fija)

VARIABLE	Importancia	%
Lengua materna	0,058	14,9%
Nivel de estudio	0,052	13,4%
Nivel de pobreza	0,046	11,9%
N.º de habitaciones	0,037	9,5%
Edad del jefe del hogar	0,037	9,5%
Tiempo en la vivienda	0,034	8,8%
Condición de la vivienda	0,029	7,5%
Área sociodemográfica	0,026	6,7%
Departamento	0,023	5,9%
Nivel socioeconómico	0,018	4,6%
N.º de personas	0,018	4,6%
Sexo	0,010	2,6%

Fuente: *Elaboración Propia*

CONCLUSIONES

Los factores de mayor importancia en el uso de las TIC en los hogares del Perú, son: la lengua materna, el nivel de estudio del jefe del hogar y el área sociodemográfica donde se encuentra ubicada la vivienda. Estos resultados fueron obtenidos a partir del modelo de minería de datos *adaboost,* con el cual se obtuvo una exactitud en la clasificación de hogares del 79%.

Respecto al uso del Internet en los hogares del Perú, se obtuvo que los factores sociodemográficos de mayor importancia para su uso son: la lengua materna y el nivel de estudio del jefe del hogar, resultados que fueron obtenidos a partir del modelo *adaboost*, con el cual se obtuvo una exactitud en la clasificación de hogares del 90%.

En relación con el uso de la tv de paga en los hogares del Perú, se obtuvo que los factores sociodemográficos de mayor importancia para su uso son: el nivel de estudio y la lengua materna del jefe del hogar, además del área sociodemográfica donde se encuentra ubicada la vivienda, resultados obtenidos a partir del modelo *adaboost,* con el cual se obtuvo una exactitud en la clasificación de hogares del 78%.

Respecto al uso de la telefonía fija en los hogares del Perú, se obtuvo que los factores sociodemográficos de mayor importancia para su uso son: la lengua materna, el nivel de estudio del jefe del hogar y el nivel de pobreza del hogar, resultados obtenidos con el modelo *adaboost*, el cual obtuvo una exactitud en la clasificación de hogares del 87%.

Respecto a los patrones de uso de las TIC en los hogares del Perú, se puede decir que son aquellos hogares con jefes de familia que tienen como lengua materna el castellano, niveles de educación superior (ya sea técnica o universitaria), además de encontrarse ubicados en áreas urbanas del país.

RECOMENDACIONES

Enfocándonos en los modelos de minería de datos utilizados en esta investigación: *random forest, adaboost* y árboles de clasificación; se comprueba que son apropiados para investigaciones relacionados al uso de las TIC, puesto que con los 3 modelos se obtuvieron niveles de exactitud en la clasificación de hogares superiores al 70%, tanto para la identificación de los factores de mayor importancia en el uso de las TIC, como para el descubrimiento de los patrones de uso de los mismos. Es importante resaltar que para la identificación de los factores de mayor importancia en el uso de las TIC, internet, tv de paga y telefonía fija; siempre con el modelo *adaboost* se obtuvieron porcentajes de aciertos en la clasificación de hogares superiores a los obtenidos con el modelo *random forest,* por lo cual fue *adaboost* el modelo seleccionado para la presentación de resultados de la investigación, y se puede considerar apropiado para realizar estudios de clasificación en temas tecnológicos.

El factor más importante en el uso de las TIC, según los resultados obtenidos en esta investigación, es la lengua materna del jefe del hogar; según estadísticas del INEI, del último Censo Nacional del 2007, en nuestro país cerca del 15% de la población sólo habla lenguas nativas como el Quechua y Aymara, lo que equivale a cerca de 5 millones de peruanos que podrían encontrarse limitados a beneficiarse de tecnologías como internet y tv de paga, que en la actualidad presentan casi la totalidad de sus contenidos en idioma español. Es un gran reto para las empresas de

difusión tecnológica y el mismo Estado, desarrollar contenidos educativos y sociales para peruanos Quechua hablantes y de otras lenguas nativas.

La gran importancia que tiene el nivel educativo en el uso de las tecnologías del hogar, como es el caso del internet, nos muestra las grandes brechas sociales que existen hoy en día en nuestro país, donde estudiar una carrera de educación superior depende en muchos casos de la disponibilidad económica de las personas. Como hemos visto en los resultados, en hogares con jefes de familia con niveles educativos bajos, el uso de Internet es de menor proporción, lo que limita principalmente a los más jóvenes, que terminan perdiéndose de un "mundo" de oportunidades y de información (Capacitaciones virtuales, acceso a bibliografía online, conocimiento de idiomas, etc.), que podrían tener a su alcance y que podría significarles mejoras sustanciales, tanto a nivel de conocimiento, social y económico. Por lo cual, siendo Perú un país en vías de progreso y desarrollo, tiene el gran reto de incrementar el acceso a Internet en la población como medio de convertir a la sociedad peruana en una "Sociedad de la Información", que conozca las ventajas del uso de Internet y aproveche las oportunidades que ofrece como herramienta de desarrollo personal, social y económico.

REFERENCIAS BIBLIOGRÁFICAS

Aler, R. (2014) *Evaluación y aprendizaje dependiente de la distribución y el coste*. Departamento de Ingeniería Informática-Universidad Carlos III de Madrid.

http://ocw.uc3m.es/ingenieria-informatica/analisis-dedatos/transparencias.pdf

Alfaro, E., Gamez, M. and Garcia, N. (2013): "adabag: An R Package for Classification with Boosting and Bagging". Journal of Statistical Software, Vol 54, 2, pp. 1–35.

Bailly, K. and Milgram, M. (2009). Boosting feature selection for neural network based regression. *Neural Networks,* 22(5-6):748–756.

Barrantes, R. (2005), *Análisis de la demanda por TICS: ¿Qué es y cómo medir la pobreza digital?* Instituto de Estudios Peruanos, Lima, Perú.

Ballestero, F. (2002). *La brecha digital. El riesgo de la exclusión en la Sociedad de la Información*. Biblioteca Fundación AUNA.

Banco Mundial (2002), *Tecnologías de la Información y de las Comunicaciones: Estrategia del Grupo Banco Mundial.*

Biau, G. (2012) *Analysis of a random forests model*. The Annals of Statistics.

http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Breiman, L. (2001), *Random forests, Machine Learning.* The Annals of Statistics*, 5-32.*

Carbonell, J. & Siekkman, J. (2007) *Machine Learning and Data Mining in Pattern Recognition.* Springer, USA.

Castells, M. (2014) *El impacto de Internet en la Sociedad: Una perspectiva global*, 10-13.

https://www.bbvaopenmind.com/wpcontent/uploads/2014/03/BBVA-Comunicaci%C3%B3n-CulturaManuel-Castells-El-impacto-de-internet-en-la sociedad-una-perspectiva-global.pdf

Comisión Económica para América Latina y El Caribe (2003). *Declaración de Bávaro*. En: *Conferencia Ministerial Regional Preparatoria de América Latina y el Caribe para la*

Cumbre Mundial sobre la Sociedad de la Información. (29-31 de enero de 2003: Bávaro, Punta Cana, República Dominicana).

http://www.eclac.cl/prensa/noticias/noticias/9/11719/Bavarofinalesp.pdf

Criminisi A., Shotton J. & Konukoglu E. (2011) *Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised*

Cumbre mundial sobre la Sociedad de la Información (2003). *Declaración de Principios*: *10-12 de diciembre de 2003*, 9-12.

http://www.itu.int/wsis/docs/geneva/official/dop-es.html

Dalgaard, P. (2008), *Introductory Statistics with R, 2nd ed., Statistics and Computing,* Springer, USA.

De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, 83, 1105–1117.

Freund, Y. & Schapire R. (1995). *A decision-theoretic generalization of online learning and an application to boosting.*

Guerrero, D. & Quinde, M. (2011). Las TIC en el Perú desde el desarrollo sostenible: una propuesta para las zonas rurales. *HUESCA: AEIPRO,* 1665-1668.

Hastie, T., Tibshirani & Friedman J. (2001).*The Elements of Statistical Learning Springer*.

Ho, T. & Hull, J. (1995) Decision combination in multiple classifier systems. *IEEE Transactions PAMI;* 16(1): 66–75

Hultstrom, K. (2013). *Image based Wheel detection using Random Forest Classification*

http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=3457767&fileOId=3459875

Instituto Nacional de Estadística e Informática. (2016). *Las Tecnologías de Información y Comunicación en los hogares*.

https://www.inei.gob.pe/media/MenuRecursivo/boletines/informe-tecnico_tecnologias-informacion-ene-feb-mar2016.pdf

Jhonson Cornell University, World Economic Forum (2015). *The Global Information Technology Report 2015.*

http://www3.weforum.org/docs/WEF_GITR2015.pdf

Kecman, V. (2001).*Learning and Soft Computing*. The Mit Press.

Montillo, A. (2009). *Random Forest, Guest Lecture.* University of Pennsylvania.

http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf

Moreno, M., Quintales, L., García F. & Martin, M. (2002) *Obtención y Validación de Modelos de Estimación de Software Mediante Técnicas de Minería de Datos.* Revista Colombiana de Computación - RCC, Vol 3, No 1.

Observatorio para la Sociedad de la Información en Latinoamérica y El Caribe (2008). *Características de los hogares con TIC en América Latina y El Caribe.*

Organismo Supervisor de Inversión Privada en Telecomunicaciones (2015). *Los servicios de telecomunicaciones en los hogares peruanos.*

https://www.osiptel.gob.pe/repositorioaps/data/1/1/1/par/erestel-2014-servicios-telecomunicaciones-hogares/ERESTEL%202012-2014.pdf

Pereira, N. (2014). *Identificación de clientes con patrones de consumo eléctrico fraudulento.*

Pérez, P. y Rodríguez, A. (2015). *El ejercicio de medir la pobreza en el Perú.* Lima, Perú.

Pérez, C. y Santín, D. (2007). *Minería de Datos: Técnicas y Herramientas*. Madrid: Ediciones Paraninfo, S.A.

Schapire, R., Freund Y., Bartlett P. & Lee W. (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics,* 26(5):1651-1686.

Sociedad Nacional de Industrias (2016). *Informe Global de la Tecnología de la Información 2016.*

http://www.cdi.org.pe/InformeGlobaldeInformacion/GITR2016.html

Stager, M. & Núñez, J. (2015) Uso de internet en Chile: la otra brecha que nos divide. *País Digital*, 39-41.

http://paisdigital.org/wpcontent/uploads/2015/07/Brecha-Digital-Internet-Estudio-Pa%C3%ADs-Digital CASEN.pdf

Swiss Federal Institute of Technology (2012).*Applied Multivariate Statistics-Spring*.

https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v10.2.pdf

Timofeev, R. (2004). Classification and regression trees (cart) theory and applications. *Humboldt University Berlin,* 4-7.

Williams, G. (2011) *Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discover.* Springer.

Wu, H. (2011). *Offline and online AdaBoost for detecting anatomic structure.* Universidad de Arizona.

https://repository.asu.edu/attachments/56917/content/Wu_asu_0010N_10863.pdf

ANEXOS

Anexos: 1. Matriz de Consistencia

PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLES	METODOLOGÍA
1. Problema Principal ¿Cuáles son los factores asociados al uso de las TIC (Internet, Telefonía fija y Televisión de paga) de los hogares en Perú en el año 2014?	**1. Objetivo General** Identificar los factores asociados al uso de las TIC (Internet, Telefonía fija y Televisión de paga) de los hogares en Perú en el año 2014.	**1. Hipótesis General** Los factores asociados en el uso de las TIC (internet, telefonía fija y televisión de paga) en los hogares del Perú en 2014 son: la lengua materna y el nivel de estudio del jefe del hogar, además del área sociodemográfica donde se ubica la vivienda.	**1. Variables independientes** Área Sociodemográfica Nivel Socioeconómico del Hogar Nivel de pobreza del hogar Condición de la Vivienda Número de miembros del Hogar Número de habitaciones de la vivienda Departamento Edad del Jefe del Hogar Tiempo de ocupación en la vivienda Nivel Educativo del Jefe del Hogar Lengua Materna del Jefe del Hogar Sexo del Jefe del hogar **2. Variables dependientes** Uso de las TIC Uso del Internet Uso de la Telefonía Fija Uso de TV de Paga.	**1. Tipo de Investigación** Aplicada, cuantitativa, descriptiva, correlacional y explicativa. **2. Diseño de la Investigación** No experimental y Transversal **3. Métodos a utilizar** a. Técnicas estadisticas exploratorias y descriptivas para la depuración, recodificación, tratamiento y análisis de los datos. b. Modelos de clasificación supervisada de Minería de datos: Random Forest, AdaBoost y Árboles de Clasificación. **4. Materiales o Instrumentos** La Base de Datos para esta investigación corresponde a la ENCUESTA RESIDENCIAL DE SERVICIOS DE TELECOMUNICACIONES ERESTEL 2014, realizada por OSIPTEL. **5. Descripción de la Muestra** a.De tipo probabilística, multietápica, estratificada, por conglomerados estratificados implícitamente por nivel socio económico y de selección sistemática d. Tamaño de Muestra:14 842 hogares.
2. Problemas Específicos	**2. Objetivos Específicos**	**2. Hipótesis Específicas**		
a. ¿Cuáles son los factores asociados al uso de Internet de los hogares en Perú en el año 2014?	**a**.Determinar los factores asociados al uso de Internet de los hogares en Perú en el año 2014.	**a.** Los factores asociados al uso de internet en los hogares del Perú en 2014 son: la lengua materna y el nivel de estudio del jefe del hogar.		
b. ¿Cuáles son los factores asociados al uso de la tv de paga de los hogares en Perú en el año 2014?	**b.** Describir los factores asociados al uso de la tv de paga de los hogares en Perú en el año 2014.	**b.** Los factores asociados al uso de la telefonía fija en los hogares del Perú en 2014 son: el nivel de estudio y la lengua materna del jefe del hogar, además del área sociodemográfica donde se encuentra ubicada la vivienda.		
c. ¿Cuáles son los factores asociados al uso de la telefonía fija de los hogares en Perú en el año 2014?	**c**.Identificar los factores asociados al uso de la telefonía fija de los hogares en Perú en el año 2014.	**c.** Los factores asociados al uso de la tv de paga en los hogares del Perú en 2014 son: la lengua materna y el nivel de estudio del jefe del hogar, además del nivel de pobreza del hogar.		
d. ¿Cuáles son los patrones de consumo de las TIC en los hogares en Perú en el año 2014?	**d.** Especificar los patrones de consumo de las TIC en los hogares del Perú en el 2014.	**d.** Los hogares del Perú que más hacen uso de las TIC son aquellas con jefes de familia que tienen como lengua materna el castellano, niveles de educación superior (ya sea técnica o universitaria), además de contar con viviendas que se encuentran ubicadas en áreas urbanas.		

Anexos: 2. Matriz de operacionalización de variables

VARIABLE	DEFINICIÓN	TIPO / ESCALA	CATEGORÍA
AREA SOCIODEMOGRÁFICA*	Caracterización realizada a una determinada zona geográfica en base a sus condiciones poblacionales.	Cualitativa Nominal Dicotómica	Urbana, Rural
NIVEL SOCIOECONOMICO DEL HOGAR*	Medida económica y sociológica combinada de la preparación laboral familiar en relación a otras familias, basada en sus ingresos, educación, y empleo.	Cualitativa Ordinal Politómica	A,B,C,D,E
NIVEL DE POBREZA*	Categorización que se realiza a partir del valor monetario de una canasta de bienes y servicios que cumple las necesidades básicas de un hogar.	Cualitativa Ordinal Politómica	No Pobre, Pobreza No Extrema, Pobreza Extrema
CONDICION DE LA VIVIENDA	Condición habitacional del hogar en la vivienda que ocupa.	Cualitativa Ordinal Dicotómica	Alquilada, Propia
NUMERO DE MIEMBROS DEL HOGAR	Cantidad de personas que conforman de manera continua el hogar.	Cualitativa Ordinal Politómica	De 1 a 2, De 3 a 4, De 5 a 6, De 7 a más personas
NUMERO DE HABITACIONES DE LA VIVIENDA	Cantidad de habitaciones que posee la vivienda, considerando como habitaciones: dormitorio, sala, cocina, comedor, baño.	Cualitativa Ordinal Politómica	De 1 a 2, De 3 a 4, De 5 a 6, De 7 a más habitaciones

LENGUA MATERNA DEL JEFE DEL HOGAR	Referido al primer idioma aprendido por el jefe del hogar.	Cualitativa Nominal Politómica	Castellano, Quechua, Aymara, Otros
SEXO DEL JEFE DEL HOGAR	Referido al género del jefe del hogar.	Cualitativa Nominal Dicotómica	Masculino, Femenino
DEPARTAMENTO	División política-geográfica principal que tiene el país a lo largo de su territorio.	Cualitativa Nominal Politómica	Todos los departamentos del Perú
EDAD DEL JEFE DEL HOGAR	Referido a los años que tiene el jefe de hogar.	Cualitativa Ordinal Politómica	Generación Z (15 a 20 años), Millennials (21 a 34 años), Generación X (35 a 49 años), Boomers (50 a 64 años), Generación Silenciosa (65 a más).
TIEMPO DE OCUPACIÓN EN LA VIVIENDA	Referido al tiempo que habita la familia en la vivienda.	Cualitativa Ordinal Politómica	Menos de 1 año, De 1 a menos de 5 años, De 5 a menos de 10 años, De 10 a menos de 15 años, De 15 años a más
NIVEL EDUCATIVO DEL JEFE DEL HOGAR	Referido al último nivel académico alcanzado por el jefe del hogar.	Cualitativa Ordinal Politómica	Sin Nivel, Primaria, Secundaria, Sup. Técnica, Sup. Universitaria
USO DE LAS TICS EN EL HOGAR	Hogar que utiliza los servicios de internet de conexión fija, tv de paga y telefonía fija.	Cualitativa Nominal Dicotómica	Si, No
USO DE INTERNET	Hogar que utiliza el servicio de internet de conexión fija.	Cualitativa Nominal Dicotómica	Si, No

USO DE TV DE PAGA	Hogar que utiliza el servicio de tv de paga (pago por canales que no son de señal abierta)	Cualitativa Nominal Dicotómica	Si, No
USO DE LA TELEFONÍA FIJA	Hogar que utiliza el servicio de telefonía fija.	Cualitativa Nominal Dicotómica	Si, No

**Variables presentadas por Osiptel, en base a la operacionalización de variables del Instituto Nacional de Estadística e Informática en la Encuesta Nacional de Hogares (ENAHO)*

Anexos: 3. Evaluación de modelos de clasificación Random Forest y Adaboost

1. Uso de las TIC

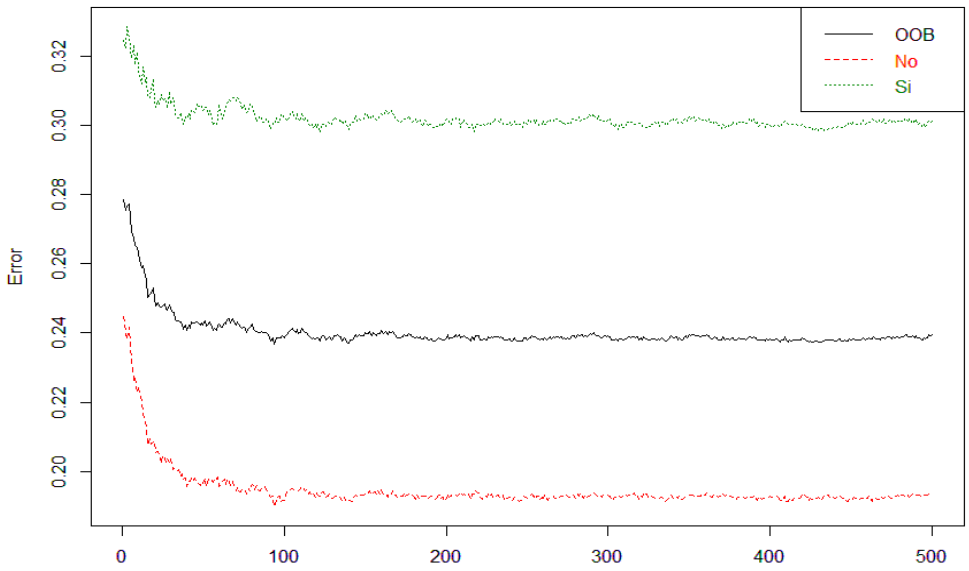
- **Random Forest**

Tabla 1: Matriz de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de las TIC)

Uso de las TIC		Predicho		Exactitud
		Sí	No	
Real	Sí	3079	1328	76,06%
	No	1159	4823	

Fuente: *Elaboración Propia*

Figura N° 1: Tasas de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de las TIC)



Fuente: *Elaboración Propia*

La exactitud obtenida en la etapa de entrenamiento es de 76,06%. De la Tabla N° 1 se observa que del total de hogares en los que se hace uso de las TIC, 3079 fueron clasificados correctamente por el modelo *random forest*. Además de la Figura

N° 1, se observa que la tasa de error de clasificación fue menor para la clasificación de hogares en los que no se hace uso de las TIC (especificidad). La curva OOB de la Figura N° 1, representa el error de clasificación de todos los hogares con el modelo *random forest.*

De la Tabla N° 2, se observa que con los datos de la muestra de prueba fueron mejor clasificados los hogares que no hacen uso de las TIC, tendencia que se apreció manera similar con la muestra de entrenamiento. La exactitud obtenida en esta etapa fue del 75%.

Tabla N° 2: Matriz de error del modelo random forest en la etapa de prueba (Variable de destino: Uso de las TIC)

Uso de las TIC		Predicho		Exactitud
		Sí	No	
Real	Sí	646	289	75,00%
	No	267	1024	

Fuente: *Elaboración Propia*

En la Tabla N° 3 se muestran las clasificaciones realizada con los datos para validar el modelo, donde también se observan estructuras de clasificación similar a los vistos con las muestras de entrenamiento y prueba. La exactitud obtenida en esta etapa fue del 77%.

Tabla N° 3: Matriz de error random forest en la etapa de validación (Variable de destino: Uso de las TIC)

Uso de las TIC		Predicho		Exactitud
		Sí	No	
Real	Sí	1892	757	77,00%
	No	694	2964	

Fuente: *Elaboración Propia*

Como se observó, los valores obtenidos con las matrices de error, tanto para la muestra de entrenamiento, prueba y validación se encuentran con niveles de exactitud similares, lo cual nos indica que el modelo *random forest* utilizado es adecuado para clasificar los hogares del Perú en los que se usan TIC.

- **Adaboost**

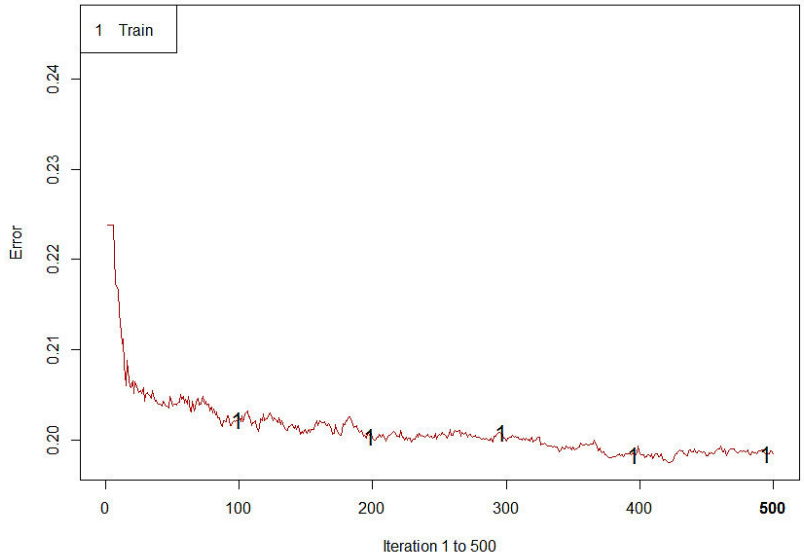
Tabla N° 4: Matriz de error del modelo adaboost en la etapa de entrenamiento (Variable de destino: Uso de las TIC)

Uso de las TIC		Predicho		Exactitud
		Sí	No	
Real	Sí	3322	1085	80,14%
	No	978	5004	

Fuente: *Elaboración Propia*

De la Tabla N° 4, se observa que, con los datos de la muestra de entrenamiento, fueron mejor clasificados los hogares que no hacen uso de las TIC. La exactitud obtenida en esta etapa fue del 80%.

Figura N° 2: Tasa de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de las TIC)



Fuente: *Elaboración Propia*

De la Tabla N° 5, se observa que, con los datos de la muestra de prueba, fueron mejor clasificados los hogares que no hacen uso de las TIC. La exactitud obtenida en esta etapa fue del 80%.

Tabla N° 5: Matriz de error del modelo adaboost en la etapa de prueba (Variable de destino: Uso de las TIC)

Uso de las TIC		Predicho		Exactitud
		Sí	No	
Real	Sí	693	242	79,00%
	No	225	1066	

Fuente: *Elaboración Propia*

En la Tabla N° 6 se muestra la exactitud en la clasificación realizada con los datos de validación que fue del 79%.

Tabla N° 6: Matriz de error del modelo adaboost en la etapa de validación (Variable de destino: Uso de las TIC)

Uso de las TIC		Predicho		Exactitud
		Sí	No	
Real	Sí	690	245	79,00%
	No	223	1068	

Fuente: *Elaboración Propia*

En la muestra de entrenamiento, de prueba y validación, podemos observar estructuras de clasificación y valores de exactitud similares, lo cual nos indica que el modelo *adaboost* utilizado es adecuado para la resolución de este tipo de problemas.

2. Uso de internet

- **Random forest**

La exactitud obtenida en la etapa de entrenamiento con el modelo *random forest* es de 84.2% para la clasificación de hogares en el Perú que hacen uso de internet.

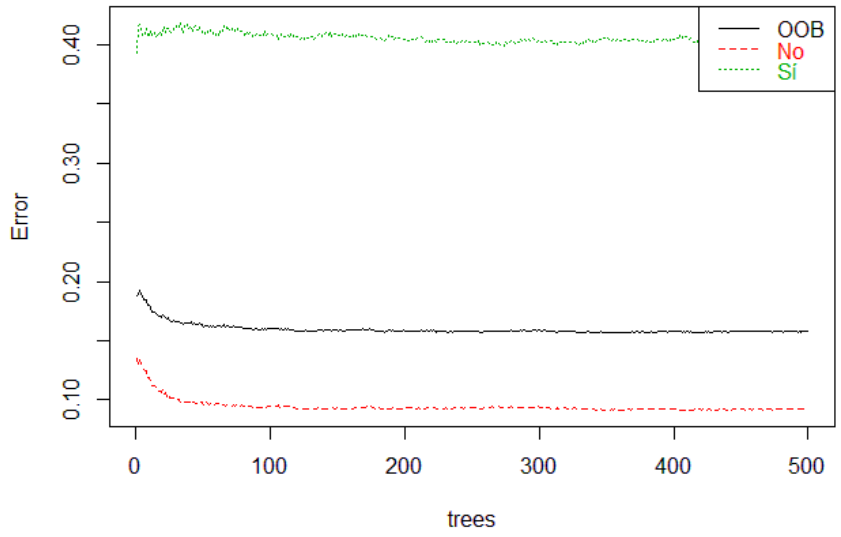
Tabla N° 7: Matriz de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de internet)

		Predicho		Exactitud
		Sí	No	
Real	Sí	1295	880	84,19%
	No	762	7452	

Fuente: *Elaboración Propia*

De la Tabla N° 7 se observa que del total de hogares en los que no se hace uso de internet, 7452 fueron clasificadas de manera correcta por el modelo *random forest* generado en la etapa de entrenamiento. Además de la Figura N° 3 se observa que la tasa de error de clasificación fue menor para la clasificación de hogares en los que no se hace uso de Internet (9.3%).

Figura N° 3: Tasas de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de internet)



Fuente: *Elaboración Propia*

Tabla N° 8: Matriz de error del modelo random forest en la etapa de prueba (Variable de destino: Uso de Internet)

		Predicho		Exactitud
		Sí	No	
Real	Sí	287	198	85,15%
	No	132	1609	

Fuente: *Elaboración Propia*

De la Tabla N° 8 se observa que con los datos de la muestra de prueba se obtuvieron valores de exactitud del 85%.

En la Tabla N° 9 se muestran los errores obtenidos en la clasificación realizada con los datos de prueba, donde también se observan proporciones de estructura de clasificación similar a los vistos con las muestras de entrenamiento y prueba, con la muestra de validación. La exactitud obtenida con la muestra de validación es del 83%.

Tabla N° 9: Matriz de error del modelo random forest en la etapa de validación (Variable de destino: Uso de Internet)

		Predicho		Exactitud
		Sí	No	
Real	Sí	289	223	83,00%
	No	156	1558	

Fuente: *Elaboración Propia*

Los valores obtenidos con las matrices de error, tanto para la muestra de entrenamiento, de prueba y de validación se encuentran con estructuras de clasificación y niveles de exactitud similares, lo cual nos confirma la idoneidad del modelo para realizar la clasificación de hogares, teniendo en consideración el Uso de Internet como variable de destino.

- **Adaboost**

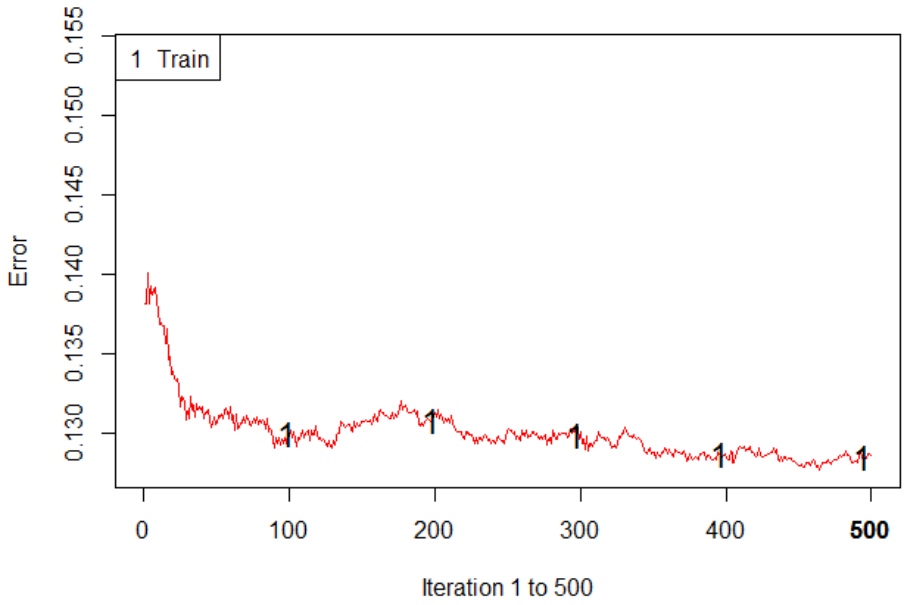
La exactitud obtenida durante la etapa de entrenamiento fue de 87.1% para la clasificación del uso de internet en los hogares del Perú mediante el modelo de minería de datos *adaboost*.

Tabla N° 10: Matriz de error del modelo adaboost en la etapa de entrenamiento (Variable de destino: Uso de Internet)

		Predicho		Exactitud
		Sí	No	
Real	Sí	1394	781	87,14%
	No	555	7659	

Fuente: *Elaboración Propia*

Figura N° 4: Tasa de error del modelo adaboost en la etapa de entrenamiento (Variable de destino: Uso de Internet)



Fuente: *Elaboración Propia*

De la Tabla N° 10 se observa que del total de hogares en los que no se hace uso de las TIC, 7659 fueron clasificadas de manera correcta por el modelo *adaboost* generado. Además de la Figura N° 4, se observa como la tasa de error total de clasificación del modelo *adaboost* se reduce conforme se incrementa el número de árboles al modelo.

De la Tabla N° 11, se observa que con los datos de la muestra de prueba el nivel de exactitud fue del 86%.

Tabla N° 11: Matriz de error adaboost (Muestra de prueba)

Variable de destino: Uso de Internet

		Predicho		Exactitud
		Sí	No	
Real	Sí	287	198	86,14%
	No	110	1631	

Fuente: *Elaboración Propia*

En la Tabla N° 12 se muestran las estructuras de clasificación y la exactitud. Con la muestra de validación, el nivel de exactitud es del 85%.

Tabla N° 12: Matriz de error adaboost (Muestra de validación)

Variable de destino: Uso de Internet

		Predicho		Exactitud
		Sí	No	
Real	Sí	289	223	85,00%
	No	111	1603	

Fuente: *Elaboración Propia*

Tanto para la muestra de entrenamiento, de prueba y de validación, encontramos estructuras de clasificación y niveles de exactitud similares, lo cual nos confirma la idoneidad del modelo para realizar la clasificación de hogares, teniendo en consideración el Uso de Internet como variable de destino.

3. Uso de TV de paga

- **Random forest**

La exactitud obtenida en la etapa de entrenamiento es de 74.34%.

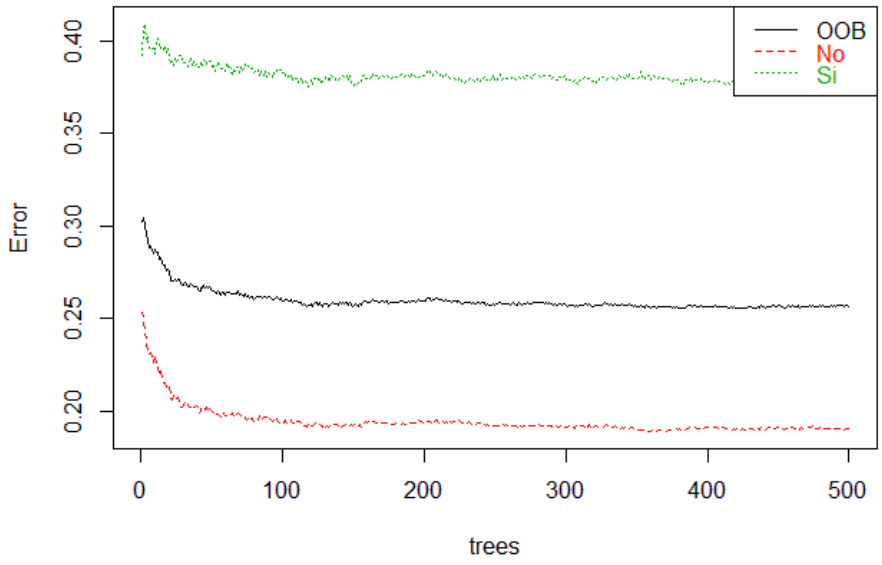
Tabla N° 13: Matriz de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de TV de paga)

		Predicho		Exactitud
		Sí	No	
Real	Sí	2264	1379	74,34%
	No	1287	5459	

Fuente: *Elaboración Propia*

Del total de hogares en los que no se hace uso de TV de paga, 5459 fueron clasificadas de manera correcta por el modelo *random forest* generado en la etapa de entrenamiento, como se observa en la Tabla N° 13. De la Figura N° 5, se observa que la tasa de error de clasificación fue menor para la clasificación de hogares en el Perú, en los que no se hace uso de TV de paga (19,1%).

Figura N° 5: Tasas de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de tv de paga)



Fuente: *Elaboración Propia*

De la Tabla N° 14, se observa que, con los datos de la muestra de prueba, se obtuvo un nivel de exactitud del 74%.

Tabla N° 14: Matriz de error random forest (Muestra de prueba) Variable de destino: Uso de TV de paga

		Predicho		Exactitud
		Sí	No	
Real	Sí	467	289	74,00%
	No	289	1180	

Fuente: *Elaboración Propia*

En la Tabla N° 15 se muestran la matriz de error y la exactitud en la clasificación realizada con los datos de validación, donde también se observan niveles similares a los vistos con las muestras de entrenamiento y prueba. La exactitud obtenida con la muestra de validación es del 75%.

Tabla N° 15: Matriz de error random forest (Muestra de validación)

Variable de destino: Uso de TV de paga

		Predicho		Exactitud
		Sí	No	
Real	Sí	472	292	74,75%
	No	270	1192	

Fuente: *Elaboración Propia*

Los valores obtenidos de exactitud tanto para la muestra de entrenamiento, de prueba y de validación del modelo *random forest,* se encuentran con niveles similares, esto nos confirma la idoneidad del modelo para realizar la clasificación de hogares del Perú teniendo en consideración el Uso de Internet como variable de destino.

- **Adaboost**

La exactitud obtenida durante la etapa de entrenamiento del modelo *adaboost* fue de 78,1%.

Tabla N° 16: Matriz de error del modelo adaboost en la etapa de entrenamiento

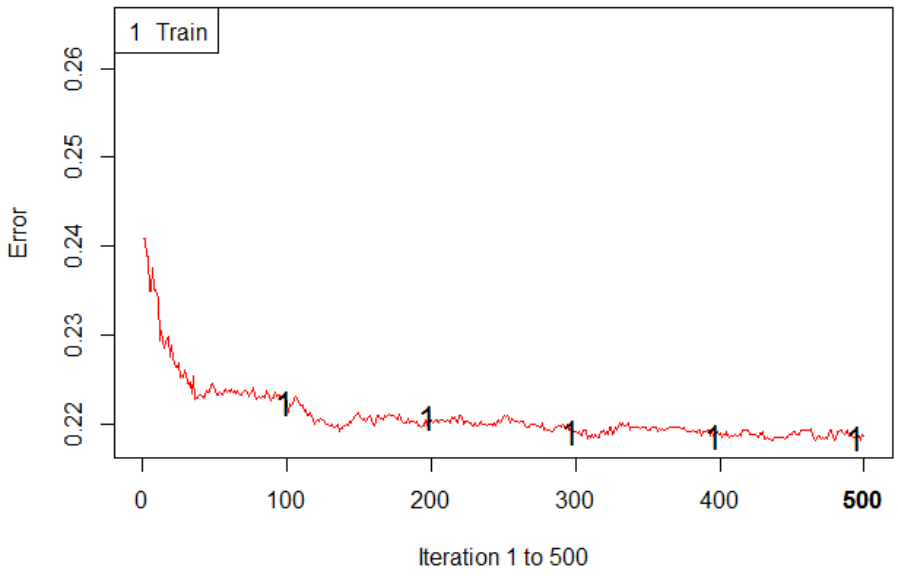
(Variable de destino: Uso de tv de paga)

		Predicho		Exactitud
		Sí	No	
Real	Sí	2287	1356	78,12%
	No	917	5829	

Fuente: *Elaboración Propia*

De la Tabla N° 16, se observa que del total de hogares en los que no se hace uso de las TIC, 5829 fueron clasificadas de manera correcta por el modelo *adaboost* generado, además de la Figura N° 6, se observa como la tasa de error de clasificación del modelo *adaboost* se reduce conforme se incrementa el número de árboles al modelo, hasta llegar al error de 21,9%.

Figura N° 6: Tasa de error del modelo adaboost en la etapa de entrenamiento (Variable de destino: Uso de TV de paga)



Fuente: *Elaboración Propia*

Tabla N° 17: Matriz de error adaboost (Muestra de prueba)

Variable de destino: Uso de TV de paga

		Predicho		Exactitud
		Sí	No	
Real	Sí	423	334	76,00%
	No	200	1269	

Fuente: *Elaboración Propia*

De la Tabla N° 17, se observa que con los datos de la muestra de prueba que fueron mejor clasificados los hogares que no hacen uso de Internet. El valor de exactitud fue del 76%.

En la Tabla N° 18 se muestran los valores obtenidos en la clasificación realizada con los datos de validación, la exactitud obtenida con la muestra de validación es del 78%.

Tabla N° 18: Matriz de error adaboost (Muestra de validación)

Variable de destino: Uso de TV de paga

		Predicho		Exactitud
		Sí	No	
Real	Sí	490	267	78,00%
	No	223	1246	

Fuente: *Elaboración Propia*

Las matrices de error y la exactitud, tanto para la muestra de entrenamiento, de prueba y de validación, se encuentran con estructuras de clasificación y niveles similares, lo cual nos confirma la idoneidad del modelo para realizar la clasificación de hogares, teniendo en consideración el uso de tv de paga como variable de destino.

4. Uso de telefonía fija

- **Random Forest**

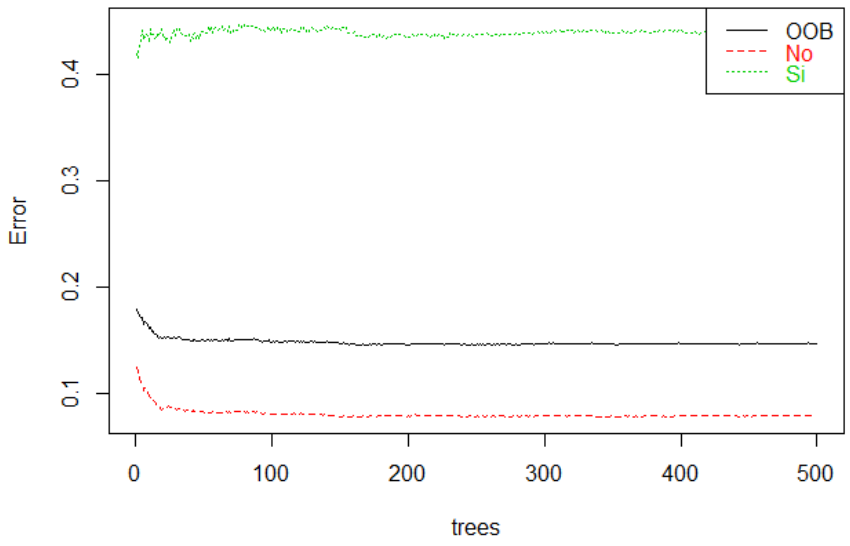
La exactitud obtenida en la etapa de entrenamiento es de 85.33% con el modelo *random forest* para la clasificación de los hogares del Perú que hacen uso de la telefonía fija. Del total de hogares en los que no se hace uso de telefonía fija, 7770 fueron clasificadas de manera correcta por el modelo *random forest* generado en la etapa de entrenamiento, como se observa en la Tabla N° 19.

Tabla N° 19: Matriz de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de telefonía fija)

		Predicho		Exactitud
		Sí	No	
Real	Sí	1095	860	85,33%
	No	664	7770	

Fuente: *Elaboración Propia*

Figura N° 7: Tasas de error del modelo random forest en la etapa de entrenamiento (Variable de destino: Uso de telefonía fija)



Fuente: *Elaboración Propia*

En la Figura N° 7, se observa las tasas de error de clasificación del modelo *random forest* en la etapa de entrenamiento, considerando como variable de destino el uso de la telefonía fija.

Tabla N° 20: Matriz de error random forest (Muestra de prueba)

Variable de destino: Uso de telefonía fija

		Predicho		Exactitud
		Sí	No	
Real	Sí	312	200	85,00%
	No	133	1580	

Fuente: *Elaboración Propia*

En la Tabla N° 21 se muestran los errores obtenidos en la clasificación realizada con los datos de validación, donde también se observa estructuras de clasificación similar a los vistos con las muestras de entrenamiento y prueba (Tabla N° 20).

Tabla N° 21: Matriz de error random forest (Muestra de validación)

Variable de destino: Uso de telefonía fija

		Predicho		Exactitud
		Sí	No	
Real	Sí	869	505	86,00%
	No	266	4667	

Fuente: *Elaboración Propia*

Las matrices de error, tanto para la muestra de entrenamiento, de prueba y de validación se encuentran con estructuras de clasificación y valores de exactitud, similares, lo cual nos confirma la idoneidad del modelo para realizar la clasificación de hogares, teniendo en consideración el uso de la telefonía fija como variable de destino.

- **Adaboost**

La exactitud obtenida durante la etapa de entrenamiento del modelo *adaboost* para la clasificación de hogares en base al uso de la telefonía fija es de 88.72%.

Tabla N° 22: Matriz de error del modelo adaboost en la etapa de entrenamiento (Variable de destino: Uso de telefonía fija)

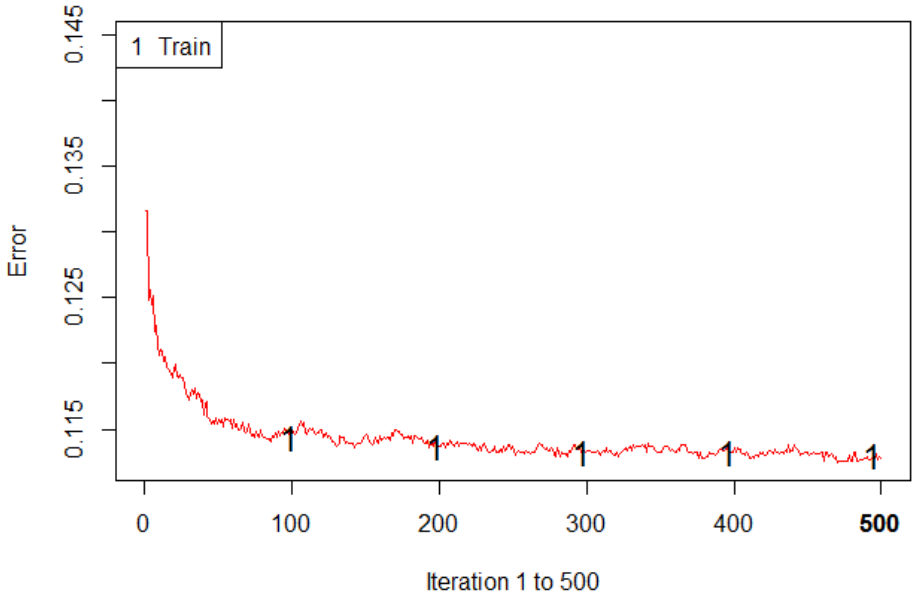
		Predicho		Exactitud
		Sí	No	
Real	Sí	1161	794	88,72%
	No	378	8056	

Fuente: *Elaboración Propia*

De la Tabla N° 22 se observa que del total de hogares en los que no se hace uso de la telefonía fija, 8056 fueron clasificadas de manera correcta por el modelo *adaboost* generado, mientras que para el caso de los hogares en los que si se hace uso de esta tecnología, 1161 fueron clasificados correctamente. En la Figura N° 8, se observan las tasas de error de clasificación del modelo *adaboost.*

Figura N° 8: Tasa de error del modelo adaboost en la etapa de entrenamiento

(Variable de destino: Uso de telefonía fija)



Fuente: *Elaboración Propia*

De la Tabla N° 23, se observa que con los datos de la muestra de prueba que la exactitud es del 88%.

Tabla N° 23: Matriz de error adaboost (Muestra de prueba)

Variable de destino: Uso de telefonía fija

		Predicho		Exactitud
		Sí	No	
Real	Sí	245	156	88,00%
	No	111	1714	

Fuente: *Elaboración Propia*

Tabla N° 24: Matriz de error random adaboost (Muestra de validación)

Variable de destino: Uso de telefonía fija

		Predicho		Exactitud
		Sí	No	
Real	Sí	328	197	88,24%
	No	63	1638	

Fuente: *Elaboración Propia*

.

Las matrices de error, tanto para la muestra de entrenamiento, de prueba y de validación se encuentran con estructuras de clasificación y valores de exactitud, similares, lo cual nos confirma la idoneidad del modelo para realizar la clasificación de hogares, teniendo en consideración el uso de la telefonía fija como variable de destino.

Anexos: 4. Clasificación de hogares del Perú según uso de las TIC mediante el modelo de árboles de clasificación

Se utilizó el modelo de árbol de clasificación CART. Para la construcción de este modelo se consideró como variable dependiente, el uso de las TIC, y como variables independientes a los 12 predictores (variables independientes) mostrados en la matriz de operacionalización (Anexo N° 2), relacionadas a las características del hogar y al jefe del hogar como variables de entrada. De la Tabla N° 25 se observa que durante la muestra de entrenamiento el modelo de árbol de clasificación CART, tuvo una exactitud del 77,14%, para esta etapa se trabajó con el 70% de los datos de la muestra total. En la etapa de validación, se obtuvo una exactitud de 77,61%, similar a lo obtenido con las muestras de entrenamiento y validación (22,9% y 21,9%).

Tabla N° 25: Matriz de error del modelo de árbol de clasificación CART. Muestra de entrenamiento y validación –Variable de destino: Uso de las TIC

Muestra	Observado	Pronosticado		Exactitud	Sensibilidad	Especificidad
		Si	No			
Entrenamiento	Si	2899	1493	77,14%	66,01%	85,38%
	No	869	5073			
Validación	Si	1254	630	77,28%	66,56%	86,36%
	No	358	2266			

Fuente: *Elaboración Propia*

Anexos: 5. Cálculo metodológico del Nivel de Pobreza – Instituto Nacional de Estadística e Informática

Para el cálculo del Nivel de Pobreza en los hogares del Perú, el Instituto Nacional de Estadística e Informática del Perú (INEI), utiliza el enfoque de la pobreza monetaria, que consiste en identificar como pobre monetario a aquel hogar que vive con un gasto per cápita insuficiente para adquirir la canasta básica de consumo de alimentos y no alimentos (vivienda, educación, vestido, salud, transporte, etc.). Estos gastos del hogar incluyen no sólo las compras sino también el pago en especies, las transferencias de otros hogares y las donaciones públicas (INEI, 2014). En este caso el algoritmo para definir al hogar como pobre o no pobre es el siguiente:

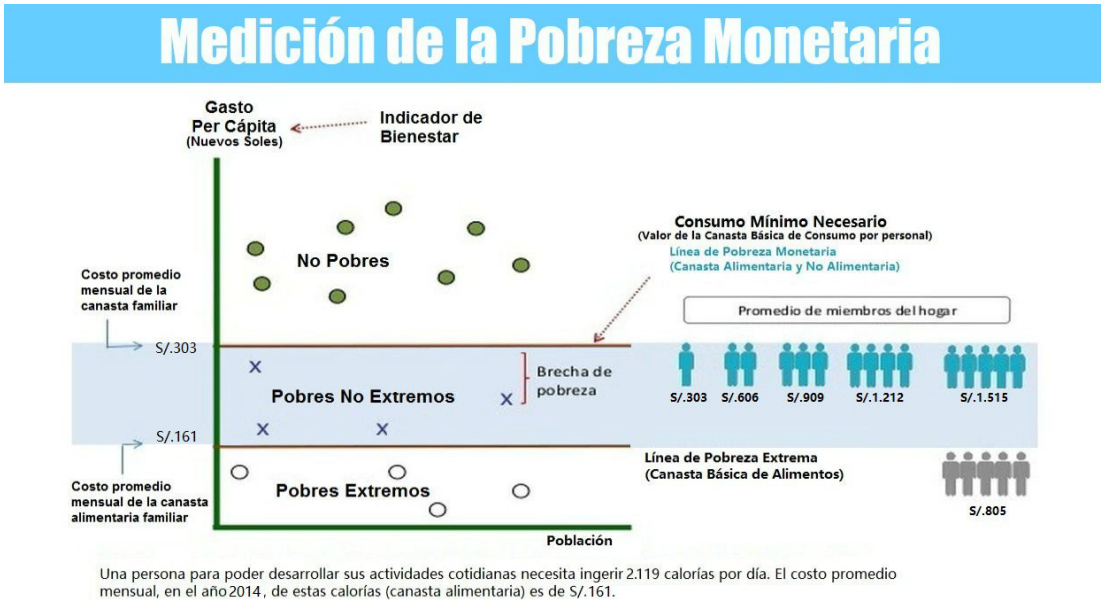
$$Y_i = \begin{cases} 1 & G_i < L_p \\ 0 & G_i \geq L_p \end{cases}$$

donde i es el hogar en cuestión, G_i es el gasto del hogar y L_p es la línea de pobreza. La estimación de la línea de pobreza entonces, parte en primer lugar, por determinar las necesidades básicas y los límites mínimos de satisfacción considerados aceptables y, en segundo lugar, consiste en valuar los mínimos aceptables de cada necesidad en términos de gasto mínimo involucrado. La sumatoria del valor se expresa en términos de ingreso total mínimo que viene a constituir la línea de pobreza.

Como se ve en el Gráfico N° 1, se establecen dos líneas de pobreza, una para la división de no pobres y pobres (S/. 303 por persona en el 2014), la cual se elabora tomando en cuenta la canasta alimentaria y no alimentaria, y otra para la división de pobres extremos y no extremos (S/. 161 por persona en el 2014), en la que sólo se tiene en cuenta la canasta de alimentos.

Gráfico N° 1: Resumen de determinación del Nivel de Pobreza en el Perú

Instituto Nacional de Estadística del Perú (INEI)



Fuente: *Adaptado de Diario El Comercio*

http://elcomercio.pe/visor/1849816/1226484-cuales-son-metodos-usados-peru-medir-pobreza-noticia